Dany LAVEAULT • Jacques GRÉGOIRE

INTRODUCTION AUX THÉORIES DES TESTS EN SCIENCES HUMAINES

MÉTHODES EN SCIENCES HUMAINES

De Boeck Université

Introduction aux théories des tests en sciences humaines

En sciences humaines, on observe une demande grandissante de développement de tests et de questionnaires. Face à cette demande, les praticiens sont souvent démunis: les bases méthodologiques et statistiques leur font défaut. Le présent ouvrage vise à combler ces lacunes. Il se veut à la fois rigoureux du point de vue des concepts et des méthodes et accessible à des non spécialistes des statistiques.

Un grand souci didactique a constamment animé les auteurs qui ont veillé à définir de manière systématique les concepts utilisés et à illustrer par de nombreux exemples les méthodes présentées. Un glossaire anglais-français et une liste des symboles apportent une aide complémentaire aux lecteurs.

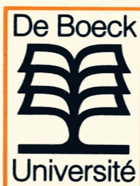
Les deux premiers chapitres sont consacrés au rappel des notions de base en statistique descriptive et inférentielle utilisées dans la suite de l'ouvrage. Les chapitres suivants abordent, pas à pas, le processus de développement d'un test: construction des items et analyse de leurs propriétés métriques, étude de la fiabilité et de la validité des scores, repérage des éventuels biais, établissement de normes et comparaison de scores.

La plus grande partie de l'ouvrage est construite sur base du modèle classique des scores. Celle-ci reste en effet la référence majeure lors du développement de tests. Les auteurs présentent cependant de manière détaillée certaines extensions de la théorie classique, comme le modèle binomial et la théorie de la généralisabilité, ainsi que les modèles récents de la réponse aux items. Le lecteur pourra ainsi se familiariser avec les développements théoriques les plus prometteurs pour la création d'outils de mesure en sciences humaines.

Dany Laveault
Professeur à la Faculté d'Éducation de l'Université d'Ottawa où il enseigne la théorie des tests, les statistiques appliquées ainsi que la méthodologie de la recherche depuis plus de quinze ans. Boursier à plusieurs reprises du Conseil de recherches en sciences humaines du Canada, il dirige en ce moment la revue Mesure et évaluation en éducation. *En 1995, il recevait le prix Benoît Poulin pour contribution exceptionnelle à la mesure en éducation.*

Jacques Grégoire
Docteur en psychologie de l'Université catholique de Louvain, il a étudié au Laboratory of Psychometric and Evaluative Research *de l'Université du Massachusetts. Il est actuellement professeur à la Faculté de Psychologie et des Sciences de l'éducation de l'Université catholique de Louvain où il enseigne la psychométrie et les méthodes d'évaluation psychologique. Il intervient régulièrement comme conseiller scientifique auprès d'organismes publics et privés pour le développement de tests psychologiques et éducatifs.*

ISBN 2-8041-2305-7



INTHTE
A 130



MÉTHODES EN SCIENCES HUMAINES

Collection dirigée par Jean-Marie De Ketele, Jean-Marie Van der Maren et Marie Duru-Bellat

COLSON J., *Le dissertoire (5e éd.)*
COSNEFROY L., *Méthodes de travail et démarches de pensée*
CRÊTE J. et IMBEAU L. M., *Comprendre et communiquer la science*
DE KETELE J.-M. et ROEGIERS X., *Méthodologie du recueil d'informations (3e éd.)*
JUCQUOIS G., *Rédiger, présenter, composer (2e éd.)*
LAVEAULT D. et GRÉGOIRE J., *Introduction aux théories des tests en sciences humaines*
LENOBLE - PINSON M., *La rédaction scientifique*
LESSARD-HÉBERT M., GOYETTE G., BOUTIN G., *La recherche qualitative. Fondements et pratiques*
MACE G., *Guide d'un projet de recherche*
PIRET A., NIZET J. et BOURGEOIS E., *L'analyse structurale*
THIRY P., *Notions de logique (2e éd.)*
VAN DER MAREN J.-M., *Méthodes de recherche pour l'éducation (2e éd.)*

Dany LAVEAULT • Jacques GRÉGOIRE

INTRODUCTION AUX THÉORIES DES TESTS EN SCIENCES HUMAINES

MÉTHODES EN SCIENCES HUMAINES

De Boeck Université

Imprimé en Belgique

D 1997/0074/102

ISSN 0779-9179
ISBN 2-8041-2305-7

À Caroline,
pour le passé, le présent et l'avenir,

À Christine, Charles et Sarah,
pour leur patience.

CHAPITRE 1

CONCEPTS DE BASE POUR UNE THÉORIE DES TESTS EN SCIENCES HUMAINES

Toute discipline scientifique aspire à mesurer et à décrire de la manière la plus précise possible les phénomènes qu'elle étudie. C'est aussi le cas de la psychologie et de l'éducation, particulièrement lorsqu'il s'agit d'avoir recours à des tests pour rendre compte d'une caractéristique, d'un trait particulier chez un sujet. C'est ici qu'entrent en jeu les notions de mesure et de statistiques nécessaires au traitement et à l'analyse des données. La quantification des variables individuelles n'est cependant pas aussi simple qu'il y paraît. Les traitements que nous pouvons réaliser sur les nombres dépendent de la nature des mesures et la description des résultats doit tenir compte des diverses propriétés de ceux-ci.

Ce chapitre propose une double incursion dans le domaine des nombres : la première dans le domaine de la *mesure* et la seconde en *statistique descriptive*. Toutes deux sont nécessaires pour bien comprendre la nature des résultats numériques que nous obtenons en notant les réponses à un test. Quant à la statistique descriptive, elle permet de mieux rendre compte de la distribution des résultats. Peut-on additionner deux résultats à des tests différents ? Comment savoir si un groupe de personnes est homogène ? La distribution des résultats obtenus permet-elle de différencier facilement les individus ? Voilà autant de questions auxquelles la mesure et la statistique descriptive essaient de répondre.

Dans ce chapitre, nous nous pencherons principalement sur les meilleurs moyens de décrire une distribution de résultats. Ces notions sont essentielles avant d'aborder les chapitres suivants. Ceux et celles qui possèdent déjà de solides notions de statistiques descriptives pourront commencer leur lecture dès le chapitre 2.

1. Les types d'échelles de mesure

Le principal intérêt d'avoir recours à un système de nombres pour effectuer les mesures en psychologie et en éducation, c'est de pouvoir se servir de leurs propriétés arithmétiques. Toutefois, avant de pouvoir effectuer une quelconque opération sur les valeurs mesurées, il faut pouvoir démontrer qu'elles correspondent à une certaine réalité, bref que cette opération est valide et qu'elle est isomorphe au système de nombres utilisé. Par exemple, deux personnes ayant chacune un quotient intellectuel de 60 ne sont pas nécessairement capables de résoudre des problèmes qu'une seule personne au quotient intellectuel de 120 serait en mesure de solutionner. Dans ce cas-ci, nous ne pouvons pas prétendre que 60+60=120.

Les échelles de mesure nous permettent de déterminer quelles opérations et quelles transformations sont possibles sur les nombres. Plus l'échelle de mesure est simple, plus ces opérations sont limitées. Plus elle est complexe, plus les opérations permises sont nombreuses. Bref, en étant bien conscients des propriétés, mais aussi des limites des échelles de mesure, nous sommes mieux préparés à utiliser les propriétés des systèmes de nombre.

Prenons un exemple courant. Nous avons l'habitude dans les compétitions sportives de nommer les joueurs par leur numéro de dossard. Ces nombres ne constituent qu'un moyen pratique d'identifier un joueur : un nom serait trop long à écrire et ne pourrait être lisible de loin. Un nombre à deux chiffres peut être imprimé avec une police en gros caractères, ce qui permet de bien identifier un joueur. Ces nombres ont tout au plus une valeur *nominale*. Il ne viendrait à l'idée de personne de les additionner ou d'en calculer la moyenne. Il en va de même des numéros de carte de crédit, d'immatriculation, de sécurité sociale...

À la base de tout travail d'administration de tests, se trouve une opération de mesure. Nous employons en effet des tests pour obtenir des informations quantitatives à propos de caractéristiques ou de traits des personnes évaluées. Pour que cette quantification ait un sens, il est crucial que les caractéristiques que l'on souhaite mesurer soient définies de manière opérationnelle. Par définition opérationnelle, il faut comprendre l'ensemble des opérations qui permettent d'obtenir une valeur caractérisant de manière valide une propriété qui nous intéresse.

Lorsque nous mesurons une caractéristique ou un trait, nous supposons que cette caractéristique ou ce trait possède une certaine permanence, une certaine stabilité. Par exemple, la mesure de la température interne du corps ne serait d'aucune utilité diagnostique chez les êtres humains si, comme chez les reptiles, elle devait changer constamment. Si l'intelligence n'était pas un trait relativement stable, nous ne serions pas intéressés à la mesurer. Lorsque nous mesurons une caractéristique, nous postulons que l'opération de mesure la laisse inchangée. Mesurer un bureau n'accroît pas la longueur de celui-ci. Cependant, avec les êtres humains, une certaine prudence s'impose. Demander à quelqu'un en thérapie de prendre en note le nombre de cigarettes qu'il fume en une journée peut sensibiliser cette personne à un point tel qu'elle en vienne à changer spontanément son comportement. Parallèlement, lorsque nous administrons un questionnaire, nous supposons que le fait de répondre aux questions ne change pas la personne qui y répond. Toutefois, ce postulat n'est pas toujours réaliste : il se peut

qu'une personne apprenne en répondant à un test et qu'ainsi les questions soient réussies différemment. Il est possible qu'un test portant sur les habitudes alimentaires sensibilise une personne au point que celle-ci réponde différemment à un traitement de prévention. C'est ce que nous appelons *l'effet de l'opération de mesure.*

Généralement, nous postulons que les facteurs précédents affectent peu ou pas notre opération de mesure. Il est alors légitime de se servir de la mesure comme d'un indicateur valable d'une caractéristique ou d'un trait que nous avons défini à un niveau théorique ou conceptuel. Toutefois nos exigences concernant la mesure peuvent être fort différentes. Au minimum, nous pouvons nous contenter d'une mesure qui ne consisterait qu'à « nommer » ou à « identifier » une caractéristique particulière à partir d'un certain nombre de propriétés communes. Au maximum, nous pouvons souhaiter obtenir une mesure qui possède tous les attributs d'un système de nombres et qui nous permette d'effectuer sur ces nombres l'ensemble des opérations arithmétiques. En d'autres termes, nous pouvons choisir d'utiliser des échelles de mesure dont les propriétés sont très variées.

Stevens (1946) a identifié quatre échelles principales de mesure fréquemment utilisées en sciences humaines et en sciences physiques :

- l'échelle *nominale* ;
- l'échelle *ordinale* ;
- l'échelle *d'intervalles* ;
- l'échelle *proportionnelle* (aussi appelée *« de rapport »).*

1.1 L'ÉCHELLE NOMINALE

C'est la plus élémentaire des formes de mesure. Comme son nom l'indique, elle consiste essentiellement à *« dénommer »* les caractéristiques mesurées. Elle est donc essentiellement qualitative et permet de regrouper dans un même ensemble les observations possédant au moins une caractéristique équivalente.

Par exemple, si dans un service de soins psychologiques nous regroupons en classes ou en catégories du DSM-IV (American Psychiatric Association, 1996), les profils diagnostiques de toutes les personnes qui consultent, nous pourrons dresser un tableau de fréquences comme le suivant :

- Cyclothymie : 23
- Dépression majeure : 18
- Dysthymie : 3

En statistique, ce type de mesure se présente sous forme de fréquences d'observations appartenant à une même classe. Dans l'exemple précédent, la fréquence des patients consultant pour une dépression majeure est de 18, alors que pour une dysthymie, cette fréquence n'est que de 3.

1.2 L'ÉCHELLE ORDINALE

Cette échelle de mesure consiste à mettre en rang les observations, d'où son nom *« échelle ordinale »*. Cette échelle est très répandue en éducation et en psychologie. Par exemple, lorsque les élèves d'un groupe-classe sont mis en rang selon leur

score total ou lorsque l'on peut placer en série différentes catégories, qu'elles soient militaires (sergent, colonel, général) ou professionnelles (ingénieur junior, ingénieur senior), nous réalisons une mesure en catégories ordinales.

1.3 L'ÉCHELLE D'INTERVALLES

Dans cette échelle, il existe une unité constante de mesure de sorte que l'intervalle entre chaque valeur de l'échelle est le même. Cette échelle possède les mêmes propriétés que l'échelle ordinale mais permet en plus de considérer que les intervalles ou écarts entre les valeurs ne changent en aucun point de l'échelle.

Avec une échelle d'intervalles, il devient possible d'affirmer qu'un écart de 10 entre un score de 40 et un score de 50 à un test est équivalent à un écart de 10 entre un score de 83 et un score de 93. De telles affirmations sont parfois difficiles à soutenir avec les scores des tests que nous employons en éducation et en psychologie, mais l'usage veut que nous agissions dans nos calculs comme si c'était vraiment le cas.

L'échelle d'intervalles possède une limite importante du point de vue métrique : elle ne possède pas de point d'origine absolu ou, si l'on préfère, aucun véritable « 0 ». Obtenir « 0 » à un test d'intelligence ne signifie pas que l'on mesure le « vide » d'intelligence. Cette valeur est donc purement arbitraire, comme c'est le cas du 0 dans l'échelle de température. En degrés centigrades de l'échelle Celsius, la valeur 0 correspond au point de congélation de l'eau au niveau de la mer. Il aurait pu tout aussi bien s'agir du point d'ébullition de l'eau ou de toute autre convention. À titre d'exemple, la valeur 0 de l'échelle Farenheit ne correspond pas au point de congélation : sur cette échelle il se situe à 32. En plus d'avoir des points d'origine différents, les échelles Celsius et Farenheit possèdent une autre différence : l'unité de mesure de température est différente. Un changement d'une unité centigrade correspond à un changement de 1,8 unités à l'échelle Farenheit.

1.4 L'ÉCHELLE PROPORTIONNELLE

On retrouve dans cette échelle de mesure toutes les propriétés d'une échelle à intervalles égaux avec, en plus, un véritable point d'origine, le zéro. Rarement possible en éducation, parfois en psychologie, elle est surtout l'apanage des sciences physiques où les mesures de masse, poids, volume sont constituées d'intervalles égaux et possèdent un véritable 0. En effet, 0 litre signifie absence de volume, tout comme 0 kilogramme représente une masse nulle. Il existe une échelle proportionnelle de mesure de température : c'est l'échelle des degrés Kelvin, possédant un véritable 0 (correspondant à -273,15 °C).

Cette échelle mérite l'étiquette de proportionnelle car, du fait de son point d'origine absolu, la quantité « 80 litres » représente bien le double de « 40 litres ». Par contre, on ne peut affirmer qu'un résultat de 120 sur une échelle d'intelligence représente une intelligence deux fois supérieure à un résultat de 60. Dans ce dernier cas, nous n'avons affaire qu'à une échelle d'intervalles.

1.5 UTILITÉ ET PROPRIÉTÉS DES ÉCHELLES DE MESURE

Psychologues et éducateurs sont partagés quant à la valeur à accorder aux résultats numériques d'un test ou d'un instrument de mesure. Pour certains, ce score total est tout au plus une échelle de mesure ordinale. En attribuant le même nombre de points à chaque item d'un test, nous créons l'illusion d'une échelle d'intervalles. Mais est-ce vraiment le cas ? Par exemple, on peut se demander si une personne qui a obtenu un score d'intérêt de 40 par rapport à une autre qui a obtenu un score de 30 manifeste le même écart d'intérêt qu'une personne ayant obtenu 15 par rapport à une autre ayant reçu un 5. Pour affirmer cela, il faudrait mesurer l'intérêt sur une échelle à intervalles égaux.

Pour de nombreux praticiens, il y a cependant de nombreux avantages à utiliser les nombres, tant que nous ne perdons pas de vue que nous opérons sur des valeurs et non sur les réalités qu'elles symbolisent. Dès que nous attribuons arbitrairement un point par question réussie, nous considérons que les items ont chacun une importance égale. L'utilisation d'une échelle à intervalles égaux est alors cohérente avec cette procédure, même si elle n'est pas nécessairement conforme à la réalité sous-jacente (Lord, 1953c ; p. 751).

Le tableau 1 résume les propriétés des échelles de mesure ainsi que les transformations possibles sur ces échelles. Il est important de retenir que les propriétés d'une échelle plus simple se retrouvent à l'intérieur d'une échelle plus complexe. Par exemple, toutes les opérations et les transformations sur une échelle ordinale sont possibles à l'intérieur d'échelles d'intervalles ou d'échelles proportionnelles, mais pas à l'intérieur d'échelles nominales.

Tableau 1 – Propriétés des échelles de mesure et opérations admissibles

	Propriétés admissibles	Transformations possibles
Échelle nominale	=	Correspondance 1 à 1
Échelle ordinale	< >	Monotone
Échelle à intervalles égaux	+ – × ÷	Linéaire
Échelle proportionnelle	0	Multiplicative

Comme le mentionne le tableau 1, une *échelle nominale* ne permet qu'une seule opération : l'équivalence. Tous les éléments d'une même classe sont considérés comme équivalents et l'extension de la classe, ou « *fréquence* », est la seule statistique que l'on puisse calculer. La seule transformation possible est la correspondance terme à terme : si pour des raisons de terminologie, on préfère utiliser la catégorie diagnostique « *psychotique* » plutôt que « *schizophrène* », ou encore la catégorie « *troubles graves de comportement* » plutôt que « *inadapté socio-affectif* », la correspondance est

possible si et seulement si celle-ci s'applique à tous les éléments de l'ensemble sans exception.

L'*échelle ordinale* permet d'établir la relation « *plus grand que* » et « *plus petit que* » entre les observations. Elle permet donc d'élaborer des séries. Des transformations sont possibles sur une échelle ordinale, tant et aussi longtemps que nous préservons l'ordre : un tel type de transformation est dit monotone. À titre d'exemple, prenons la question d'attitude suivante :

Si vous pouviez disposer d'un programme informatique facile d'usage pour vous aider à évaluer vos élèves, quel service souhaiteriez-vous qu'il vous rende ? Cochez la case appropriée.
1. enregistrer et classer mes propres questions ***beaucoup*** ❑ ❑ ❑ ❑ ❑ ***pas du tout***

Dans cette échelle de Likert, il importe peu que « *beaucoup* » corresponde à 5 à 4 ou à 10. Ce qui est important, c'est que *beaucoup* corresponde à la valeur la plus élevée (ou la plus faible). Si « *beaucoup* » vaut 5, on peut attribuer aux échelons suivants les valeurs 4, 3, 2 et 1. Ceci signifie que la catégorie « *pas du tout* » se voit attribuer un point. Si l'on souhaite que cette dernière corresponde à un zéro arbitraire, on peut opter pour une échelle dont la série de valeurs correspondantes serait 4, 3, 2, 1, 0. De cette manière, le point médian de l'échelle correspond à la valeur « 2 » ce qui représente exactement à la moitié de « 4 », ce qui n'était pas le cas avec une échelle 5, 4, 3, 2, 1. Toutefois, cette précision n'est qu'apparente. On pourrait tout aussi bien justifier l'échelle 10, 7, 5, 3, 0 si nous constations que les gens ont tendance à choisir les valeurs médianes et si nous souhaitions accorder plus d'importance aux choix extrêmes. Cette transformation est toujours monotone, même si elle change les intervalles entre les catégories de réponses.

L'*échelle d'intervalles* est sans doute, avec l'échelle proportionnelle, la plus séduisante. Elle permet en effet de réaliser toutes les opérations arithmétiques sur les unités de mesure, car celles-ci sont égales. Grâce à ces opérations, il sera possible de calculer des indicateurs statistiques utiles tels que la moyenne et la variance. Lorsque l'on décide de transformer une telle échelle, il faut préserver l'égalité des intervalles et tenir compte du caractère arbitraire du point d'origine. C'est pourquoi seule une transformation linéaire est possible dans le cas d'échelles d'intervalles. La transformation linéaire prend la forme de l'équation suivante :

$$Y = aX + k \tag{1.1}$$

Une belle illustration de ce genre de transformation est la transformation des degrés Celsius en degrés Farenheit, selon l'équation suivante :

$$F = 1{,}8C + 32 \tag{1.2}$$

Dans l'équation (1.2), la valeur de la constante multiplicative a est égale à *1,8*. Elle représente le nombre d'unités de degrés Celsius dans un degré Farenheit. La constante additive k constitue une correction du point d'origine arbitraire : le point de congélation est *0* en degrés Celsius et *32* en degrés Farenheit. De telles transformations linéaires sont fort répandues en éducation et en psychologie lorsque nous désirons

transformer les résultats brutes à un test en une échelle simplifiée. Ces transformations sont discutées en détail dans le chapitre 7.

L'*échelle proportionnelle* permet d'effectuer toutes les opérations arithmétiques sur les intervalles entre les valeurs et sur les valeurs elles-mêmes. Pour transformer les valeurs d'une échelle proportionnelle, il suffit de la multiplier par une constante. En effet, puisqu'il existe un véritable 0 dans toutes les monnaies, la transformation n'a pas à tenir compte du caractère arbitraire de l'origine. Dans l'équation (1.1), $k=0$ ce qui revient à poser :

$$Y = aX \tag{1.3}$$

Ce type de transformation est « *multiplicative* ». Par exemple, si 1 $ canadien vaut 25 francs belges, on peut trouver le nombre de francs belges *(FB)* correspondant à une somme exprimée en dollars canadiens en effectuant la multiplication suivante :

$$FB = 25 \times \$CND \tag{1.4}$$

Nous effectuons également une transformation multiplicative lorsque, par exemple, les résultats d'un test exprimés sur 50 sont exprimés sur 100 en les multipliant par deux.

En résumé, il est important de connaître la précision de l'échelle de mesure employée afin de déterminer le type de transformation qu'il est possible d'effectuer sur les résultats ainsi que le type de traitement statistique, *paramétrique* ou *non paramétrique* (voir chapitre 2, section 5). Enfin, les échelles de mesure exercent également une influence sur la manière dont nous pouvons caractériser une distribution de résultats.

2. Caractéristiques d'une distribution

Lorsque nous sommes en présence d'un ensemble de résultats, que ce soit les résultats à un test, les scores d'un examen ou une série d'autres mesures de grandeur, nous cherchons habituellement à les résumer et à les représenter graphiquement de manière à saisir l'essentiel de l'information numérique. Une représentation graphique souvent utilisée est l'*histogramme de fréquences.* La figure 1 illustre un tel graphique où des scores ont été regroupés en catégories d'une étendue de 20 points. On peut constater que la distribution des résultats n'est pas parfaitement symétrique. Un grand nombre de valeurs se situent entre 40 et 60 et peu de valeurs s'en éloignent. Les valeurs ont donc tendance à se regrouper vers cette catégorie et se dispersent lentement vers les extrémités.

L'observation de cet histogramme nous permet déjà de percevoir de manière intuitive plusieurs caractéristiques essentielles d'un distribution de scores. Ces caractéristiques constituent autant de paramètres que nous allons analyser en détail dans la suite de la présente section.

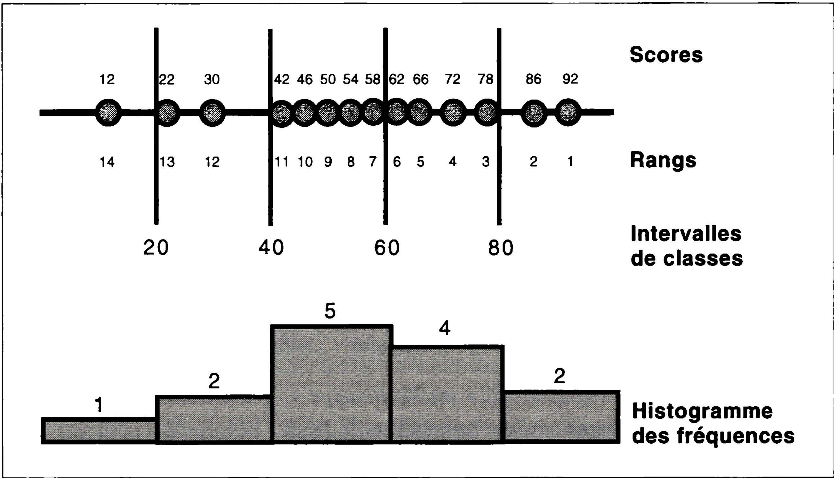


Figure 1 – Exemple d'une distribution de fréquences et d'un histogramme

2.1 VALEURS DE TENDANCE CENTRALE

Lorsque nous sommes en présence d'une série de résultats, nous souhaitons habituellement la caractériser au moyen d'indicateurs décrivant la distribution. L'indicateur le mieux connu et sans doute le plus utilisé est la *moyenne.*

La moyenne se définit comme étant l'*espérance mathématique* d'un ensemble de valeurs. C'est donc la valeur qui constitue la meilleure prédiction pour chaque valeur individuelle. En effet, si l'on fait la somme des écarts à la moyenne, l'on obtient toujours 0. On peut dès lors représenter la moyenne μ de la façon suivante :

$$\mu = E(X) \tag{1.5}$$

Par exemple, si quelqu'un avait connaissance de la moyenne des résultats avant un examen, il ferait la plus petite erreur de prédiction en attribuant ce score moyen à chacun des répondants. C'est ce qu'illustrent l'exemple suivant :

	1	2	3	4	5	Valeurs observées
–	3	3	3	3	3	– Valeurs prédites = moyenne
	-2	-1	0	+1	+2	Somme des écarts = 0

La moyenne est toutefois mieux connue par sa procédure de calcul. L'équation suivante nous indique que pour calculer une moyenne, il faut additionner chacune des valeurs et en diviser la somme par le nombre de valeurs *n*.

$$m = \frac{\sum X}{n} \tag{1.6}$$

Les deux autres valeurs de tendance centrale les plus employées sont le mode et la médiane. Observez les deux séries de valeurs suivantes :

Série A : 1 3 3 3 5

Série B : 1 2 3 4 5

Ces deux séries possèdent la même moyenne, mais dans le cas de la série A, l'un des scores apparaît beaucoup plus fréquemment. Ce score le plus fréquent d'une distribution est ce que nous appelons le *mode*. Dans la série A, le mode vaut 3. Comme toutes les valeurs ont la même fréquence dans la série B, il n'y a pas de mode.

Le calcul du mode est relativement simple. Dans une première étape, il faut calculer la fréquence de tous les scores. Le score dont la fréquence est la plus élevée constitue le mode. Par exemple, dans le cas des données de la figure 1, le mode correspond à l'intervalle de scores entre 40 et 60 (fréquence = 5). Le point milieu de cet intervalle étant 50, nous dirons que le mode de cette distribution vaut 50.

Voici deux nouvelles séries. Ces deux séries sont séparées au centre par le même score de 3.

Série A : 1 2 3 4 5

Série B : 1 2 3 4 15

Nous dirons que ces deux séries possèdent la même *médiane*. En effet, dans les deux cas, la valeur « 3 » sépare chaque série de nombres en deux moitiés égales : il y a autant de scores au-dessus qu'en dessous de 3 dans les deux séries. Par contre, la moyenne de la série B est beaucoup plus élevée. Elle tient compte non seulement de la position des nombres mais aussi de leur grandeur ou poids relatif dans la distribution. La moyenne de la série A est égale à 3, alors que celle de la série B est égale à 5. Cette différence est imputable à une seule valeur extrême, le score 15.

Pour calculer la médiane, il faut d'abord placer les données en rangs. Ensuite, il faut calculer le rang occupé par la médiane dans la distribution. Le rang de la médiane est fourni par l'équation suivante où n indique le nombre de données mises en rang :

$$rang_{méd} = \frac{1+n}{2} \quad (1.7)$$

Dans la figure 1, la médiane occupe le rang (1+14)/2, soit le rang 7,5. La médiane correspond donc au score qui se situe entre celui qui occupe le rang 7 (50) et celui qui occupe le rang 8 (54). Par intrapolation, nous prendrons le point milieu entre ces deux scores et dirons que la médiane vaut 52.

Ces propriétés différentes de la moyenne et de la médiane font que la médiane est considérée comme le *« centre de position »* alors que la moyenne est le *« centre de gravité »* d'une distribution de scores. Les propriétés particulières de ces deux valeurs de tendance centrale nous sont fort utiles lorsque nous devons apprécier le degré de symétrie d'une distribution de scores. En effet, lorsque la moyenne et la médiane

coïncident, la distribution est généralement symétrique. Par contre, lorsqu'il y a un écart entre la moyenne et la médiane, il y a asymétrie dans la distribution des résultats.

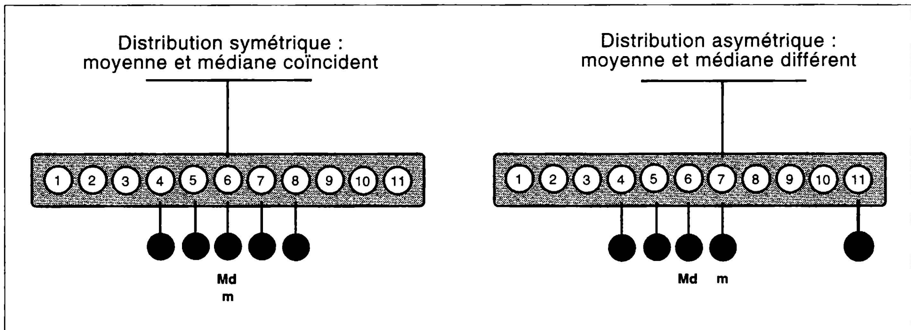


Figure 2 – Illustration d'un centre de position et d'un centre d'équilibre sur une balance à fléau

La figure 2 illustre le phénomène de la symétrie au moyen de l'exemple de deux balances. Dans la première balance, les poids sont suspendus également de part et d'autre du centre de position. Le fléau de la balance est en équilibre, car le centre de position et le centre de gravité coïncident. Dans la seconde balance, un poids est déplacé à une extrémité. Pour rétablir l'équilibre, il faut déplacer le pivot de la balance vers le centre de gravité. Tous, dans notre enfance, nous avons connu ce phénomène de la bascule où, pour jouer avec un enfant plus lourd ou plus léger, il fallait déplacer le pivot de la bascule. Moyenne et médiane d'une distribution symbolisent le même phénomène dans une distribution de scores. La moyenne est influencée par le poids relatif de chaque score, alors que le point milieu (situé à 5 dans la figure 2) n'est pas influencé par les autres valeurs.

2.2 Valeurs importantes de position

En plus des valeurs du mode, de la médiane et de la moyenne, d'autres scores occupent des positions intéressantes à l'intérieur d'une distribution. Il s'agit des *quartiles* qui divisent une distribution de scores en quatre parties égales, des *déciles* et des *centiles* qui divisent une distribution respectivement en 10 et 100 parties égales. Tout comme la médiane, ce sont des valeurs de position qui requièrent que nous placions les données en rang.

Ces valeurs de position permettent de situer rapidement une personne par rapport à un groupe de référence. Obtenir un résultat de 19/25 peut signifier plusieurs choses. Informer une personne qu'elle est la trentième de son groupe ne lui apprend rien si elle ignore combien de personnes ont été évaluées. En effet, occuper le trentième rang sur 100 est loin de représenter une performance comparable à celle qui consisterait à occuper le trentième rang sur 1000. Si, par contre, nous savons que le score 19 occupe le rang centile 82, alors nous savons que pour chaque tranche de 100 personnes évaluées, 82 obtiennent un score inférieur à 19 et 18 un score supérieur.

Le rang centile *RC* d'un score est donné par la formule suivante :

$$RC = 100 - \frac{100R - 50}{N} \qquad (1.8)$$

où *R* représente le rang du score dont on cherche le rang centile et *N* le nombre de scores de la distribution.

En guise d'exemple, calculons le rang centile du score 30 de la distribution de scores de la figure 1. Ce score occupe le douzième rang en ordre décroissant *(R = 12)* d'une série de quatorze *(N = 14)*. En substituant ces valeurs dans la formule précédente, nous trouvons :

$$RC = 100 - \frac{(100 \times 12) - 50}{14} = 16 \qquad (1.9)$$

La valeur obtenue est arrondie à l'entier le plus proche. Une valeur de 16 signifie donc que 16% des sujets ont obtenu un score inférieur à 30 et 84% un score supérieur.

La mise en rangs centiles correspond au besoin de rapporter à une échelle pratique — dans ce cas-ci de rangs — les scores d'une distribution. Ce genre de transformation est semblable à celle que nous effectuons lorsque nous ramenons un score à une échelle de pourcentage. Un score de 10 sur 15 correspond à un pourcentage de 67% alors qu'un score de 10 sur 20 correspond à 50%. Dans le cas du calcul des centiles, il ne faut pas oublier que ce n'est pas une transformation du score qui est en jeu, mais une transformation de son rang.

À partir du rang centile, il est possible de déterminer d'autres points intéressants au sein d'une distribution. Les quartiles 1, 2 et 3, par exemple, correspondent aux rangs centiles 25, 50 et 75. Les déciles 1, 2 et suivants correspondent aux centiles 10, 20 et suivants. La médiane correspond au rang centile 50 ou, si l'on préfère, au décile 5 ou encore au quartile 2.

2.3 VALEURS DE DISPERSION

Nous avons cependant besoin d'autres valeurs en plus de la tendance centrale pour définir de façon précise une distribution de scores. Observons les deux séries de scores suivantes :

Série A : 1 2 3 4 5

Série B : 2 2 3 4 4

Même si les deux séries ont la même moyenne et la même médiane, la dispersion des résultats n'est pas la même. Le moyen le plus simple de s'en rendre compte est de calculer la différence entre le maximum et le minimum de chaque série. L'écart est de 4 dans la série A et de 2 dans la série B. Pour être tout à fait rigoureux, il faudrait tenir compte de l'étendue entourant chaque valeur discrète. Le véritable minimum n'est pas 4 mais sa borne inférieure sur une échelle continue, soit 3,5. De même pour le maximum, la valeur supérieure de la borne du score 5 est 5,5. Une première valeur de dispersion d'une série de scores nous est donc donnée par l'*étendue (E)*, que nous calculons de la manière suivante pour les raisons énoncées précédemment :

$$E = (Max - Min) + 1 \quad (1.10)$$

L'étendue de la série A vaut donc 5. Toutefois, cette valeur n'est pas très précise comme indice de dispersion. Elle ne tient compte que des scores extrêmes, ce qui n'est pas très représentatif. Dans l'exemple suivant, les séries A et B ont les mêmes valeurs de tendance centrale (moyenne, médiane) et les mêmes étendues.

Série A : 1 2 3 4 5

Série B : 1 1 3 5 5

Pourtant, ces deux séries représentent des dispersions différentes. Dans la série « A », les valeurs 2 et 4 s'écartent moins de la valeur de tendance centrale que les valeurs 1 et 5. Dans la série « B », les valeurs sont plus extrêmes, bien que réparties de manière symétrique de part et d'autre de la moyenne.

En supposant que ces valeurs soient des mesures d'intervalles, pouvons-nous calculer un indice numérique de la dispersion autour de la moyenne ? La somme des écarts à la moyenne serait en apparence indiquée. Elle n'est cependant d'aucune utilité puisque, comme nous l'avons déjà démontré, cette somme vaut 0. En élevant les valeurs des écarts au carré, il est possible d'obtenir une somme non nulle car les valeurs négatives élevées au carré deviennent alors positives. En divisant cette somme des écarts au carré par le nombre total de valeurs, nous obtenons une valeur moyenne de dispersion qui n'est pas influencée par le nombre d'écarts. Cet indice de dispersion se nomme la *variance*. Le tableau 2 en fournit un exemple de calcul.

Tableau 2 – Exemple de calcul de la variance

1	2	3	4	5	Valeurs observées
– 3	3	3	3	3	– Moyenne
–2	–1	0	1	+2	Écarts à la moyenne
4	1	0	1	4	Écarts à le moyenne au carré
	+	+	+	+	
		10			Somme des écarts au carré
		$\frac{10}{5} = 2$			Moyenne des écarts au carré = Variance

L'ensemble des opérations nécessaires au calcul de la variance trouve sa traduction symbolique dans l'équation suivante :

$$s^2 = \frac{\sum (X - \bar{X})^2}{n} \quad (1.11)$$

où X représente les scores observés, $\bar{X}$ représente la moyenne des scores et n le nombre total de scores. La lettre s élevée au carré est, par convention, la lettre qui symbolise la variance. Elle nous rappelle que la variance est un indice de dispersion qui s'exprime en unités au carré. C'est pourquoi il est parfois préférable d'utiliser l'*écart*

type s qui exprime la dispersion dans le même système d'unités que la moyenne. L'écart-type n'est autre que la racine carrée de la variance :

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} \qquad (1.12)$$

En plus de l'écart-type et de la variance, il existe un autre indicateur pratique de la dispersion des résultats qui tient compte de la position des valeurs plutôt que de leur grandeur relative. Cet indice de dispersion convient particulièrement à des mesures ordinales : c'est l'*intervalle semi-interquartile.* Il s'agit en fait de calculer l'étendue entre deux positions particulièrement significatives autour de la médiane : le premier et le troisième quartile. Comme l'illustre la figure 3, l'étendue entre le troisième et le premier quartile nous donne une indication de la dispersion de 50% des valeurs autour de la médiane.

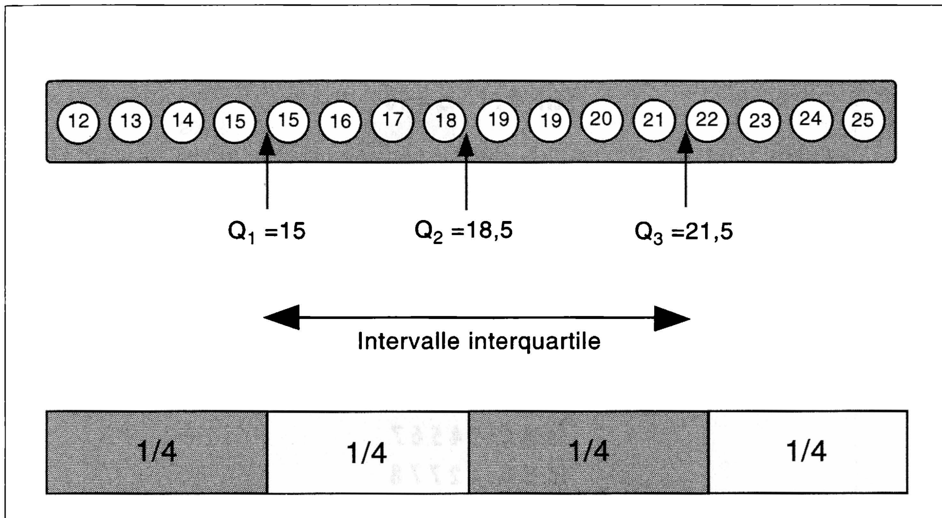


Figure 3 – Illustration de l'intervalle semi-interquartile d'une distribution

Tout comme la médiane, les quartiles ne sont pas influencés par le poids des valeurs extrêmes. Substituez 9 à 12 et 30 à 25 dans l'exemple précédent et l'intervalle interquartile ne change pas. Cette mesure est donc principalement utile pour juger de la dispersion des scores à proximité de la médiane. Toutefois, cet indice de dispersion est moins représentatif que l'écart-type puisqu'il est calculé à partir de la moitié des scores seulement. Par convention, c'est la moitié de l'étendue de l'intervalle interquartile qui sert d'indice de dispersion. L'équation décrivant la procédure de calcul de l'intervalle semi-interquartile est la suivante :

$$I = \frac{Q_3 - Q_1}{2} \qquad (1.13)$$

Dans cette dernière équation, l'intervalle semi-interquartile I est calculé en divisant par deux l'écart entre le troisième quartile Q_3 et le premier quartile Q_1. Lorsque l'écart entre le premier quartile et la médiane est différent de l'écart entre le troisième

quartile et la médiane, cela indique une accumulation des scores d'un côté ou de l'autre de la médiane.

Pour calculer la valeur de l'intervalle semi-interquartile de l'exemple de la figure 3, nous devons substituer les valeurs du troisième et du premier quartile dans l'équation (1.13). La valeur du premier quartile est 15, celle du troisième quartile est de 21,5 (valeur médiane entre 21 et 22). Nous obtenons alors le résultat suivant :

$$I = \frac{21,5 - 15}{2} = 3,25 \qquad (1.14)$$

2.4 Valeurs de symétrie

Les valeurs de tendance centrale et de dispersion nous permettent de décrire avec précision une distribution de scores. Mais là encore, d'importantes informations nous manquent pour décrire complètement la distribution des résultats. L'une de ces informations a trait à la symétrie. Observez bien les deux séries suivantes :

Série A : 1 1 5 9 9

Série B : 1 1 5 8 10

Ces deux séries de scores ont la même moyenne (5) et la même variance (13). Pourtant, la distribution des résultats est symétrique dans la série A, alors qu'elle est asymétrique dans le cas de la série B. Nous avons besoin d'un nouvel indicateur qui nous renseigne sur le degré de symétrie d'une distribution de résultats.

La procédure la plus simple pour estimer le degré d'asymétrie d'une distribution est de comparer les valeurs de la moyenne et de la médiane. Lorsque la moyenne est plus grande ou plus petite que la médiane, c'est le signe évident d'une asymétrie des résultats. Prenez en considération les deux séries de données suivantes :

Série C : 3 4 5 6 7

Série D : 1 2 7 7 8

Dans les deux séries, les moyennes sont identiques (5). Cependant, la médiane de la série D est supérieure à celle de la série C. La médiane pour C est de 5, alors que la valeur de la médiane en D égale 7. Lorsque la moyenne est inférieure à la médiane, nous parlons d'*asymétrie négative*. Dans le cas où elle est supérieure à la médiane, nous parlons d'*asymétrie positive*. Par contre, lorsque médiane et moyenne coïncident, on ne peut pas conclure qu'il y a nécessairement symétrie. Dans les séries A et B précédentes, médiane et moyenne sont égales sans pour autant que les deux distributions soient symétriques.

Une asymétrie négative est le signe d'un entassement des valeurs au-dessus de la moyenne et d'un nombre réduit de valeurs beaucoup plus petites. Une asymétrie positive est le signe d'un entassement des valeurs plus petites et d'un petit nombre de valeurs très élevées. La figure 4 représente les formes caractéristiques de chacune de ces distributions.

L'observation de la figure 4 permet de constater l'étalement des scores à l'une des extrémités de chaque distribution asymétrique. On peut donc compter sur l'observation de la distribution des résultats pour évaluer l'asymétrie d'une distribution de

scores. Cette façon de procéder demeure toutefois approximative, tout comme la comparaison des valeurs de la moyenne et de la médiane. On peut être plus précis en calculant plusieurs indices numériques d'asymétrie. L'un des ces indices met à profit l'écart entre la médiane (Md) et la moyenne dans une distribution asymétrique :

$$A = \frac{3\,(\bar{X} - Md)}{s} \tag{1.15}$$

Dans l'équation précédente, s symbolise l'écart-type et A la valeur d'asymétrie recherchée.

Une autre indice met à profit l'étalement des scores autour de la médiane. Plus la distribution est asymétrique, plus il y aura une grande différence entre l'étendue des scores Q_3 - Md et Q_1 - Md. C'est ce qu'illustre la figure 5.

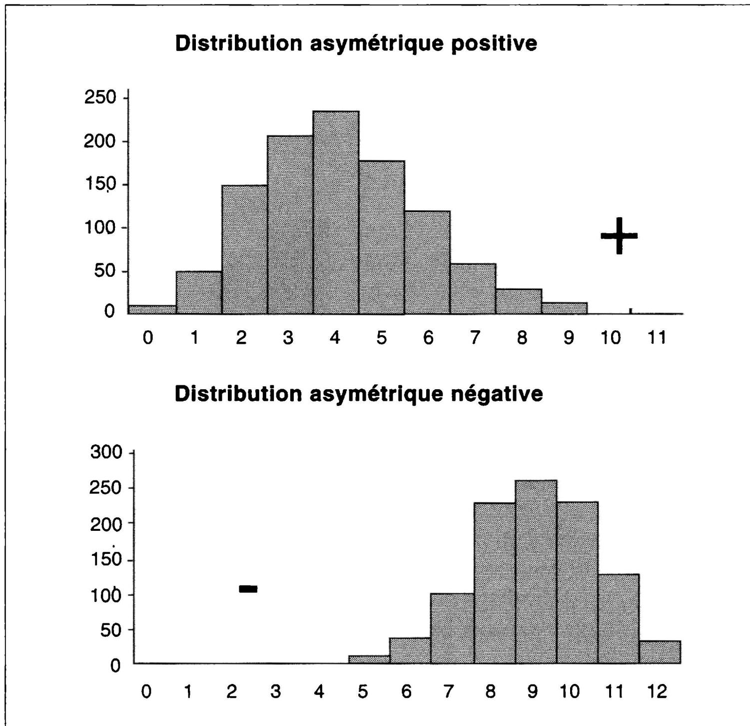


Figure 4 – Histogrammes de fréquences représentant des distributions asymétriques

On y constate un étalement des valeurs faibles et un regroupement des valeurs au-dessus de la moyenne. La distribution de fréquences de la figure 5 a été découpée exactement en quatre parties, chaque partie étant noircie par un patron différent. Comme il y a en tout 48 sujets, chaque partie grisée rend compte des résultats de 12 sujets. On constate que les 12 premiers sujets (scores inférieurs au premier quartile) ont obtenu des résultats entre 1 et 4 (écart de 4). Les 12 sujets suivants, qui ont obtenu un score entre le premier et le deuxième quartile sont beaucoup moins dispersés : leurs scores s'étendent entre 5 et 6 (écart de 2). Entre le troisième quartile et le deuxième

quartile, l'écart des scores n'est plus que de 1 car 12 des 48 sujets ont obtenu le même score de 7. Comme on peut le constater sur cette figure, une asymétrie négative va se traduire par un plus grand étalement des valeurs sous la médiane et par une concentration des valeurs au-dessus.

Ces propriétés des étendues interquartiles ont donné lieu à un autre procédé de calcul de l'asymétrie, particulièrement approprié dans le cas de données ordinales. Toutefois, ce procédé est moins précis car il ne fait pas intervenir toutes les données (80% seulement) et qu'il n'est pas sensible à la valeur relative des scores. Au lieu de limiter la mesure de l'étendue au quartile, ce procédé l'étend au centile 90 et au centile 10 de manière à inclure un plus grand nombre de valeurs. Le même raisonnement s'applique tout comme dans l'exemple de la figure 4. Il est toutefois relativement simple à calculer une fois que l'on dispose des valeurs du centile 90 (C_{90}) et du centile 10 (C_{10}) :

$$A = \frac{C_{90} + C_{10}}{2} - C_{50} \tag{1.16}$$

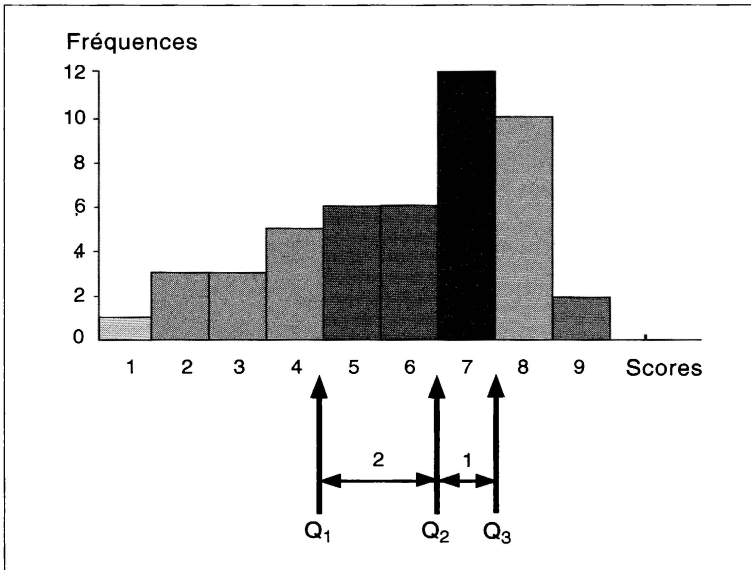


Figure 5 – L'asymétrie d'une distribution révélée par l'étalement des scores autour de la médiane

La valeur la plus rigoureuse de calcul de l'asymétrie est sans doute celle qui tient compte de la totalité des valeurs. L'asymétrie est alors obtenue de la façon suivante :

$$A = \frac{e_3}{(e_2)^{3/2}} \tag{1.17}$$

où $$e_3 = \frac{\sum (X - \bar{X})^3}{n} \text{ et } e_2 = \frac{\sum (X - \bar{X})^2}{n}$$

Cette valeur d'asymétrie vaut 0 lorsque la distribution est symétrique. Elle prend des valeurs négatives ou positives d'autant plus élevées que le degré d'asymétrie est élevé dans un sens ou l'autre de la distribution des résultats.

2.5 Valeurs de voussure de la distribution

On pourrait croire que valeurs de tendance centrale, de dispersion et d'asymétrie suffisent à caractériser une distribution. Ce serait oublier une autre caractéristique de la distribution des résultats qui nous renseigne sur le degré d'homogénéité des scores de la distribution. Visuellement, cette quatrième caractéristique se présente comme le degré de voussure plus ou moins prononcé de la distribution des résultats. Il est possible de calculer un indicateur numérique de ce degré de voussure ou d'aplatissement : la *kurtose*. Observez les deux séries de scores suivantes :

Série A : 3 3 4 5 6 7 7

Série B : 2 5 5 5 5 5 8

Ces deux séries ont mêmes moyennes, mêmes médianes, mêmes variances et elles sont toutes deux symétriques. Pourtant, elles sont manifestement différentes. Dans la série A, les valeurs sont également dispersées sur toute l'étendue des scores de la distribution. Cette étendue n'est pas aussi grande que celle de la série B, mais dans la série A, presque toutes les valeurs contribuent à la dispersion des résultats (sauf le 5). Dans la série B, le même score se répète souvent et la variance des résultats n'est imputable qu'à deux cas extrêmes.

La kurtose mesure le degré d'aplatissement d'une distribution. On en distingue trois types : les distributions leptokurtiques, élancées, concentrant un grand nombre de scores près de la moyenne, les distributions platykurtiques, évasées, se caractérisant par une répartition étendue des scores et enfin, la distribution mésokurtique, représentant une situation intermédiaire entre les deux précédentes. La figure 6 illustre une distribution de chaque type.

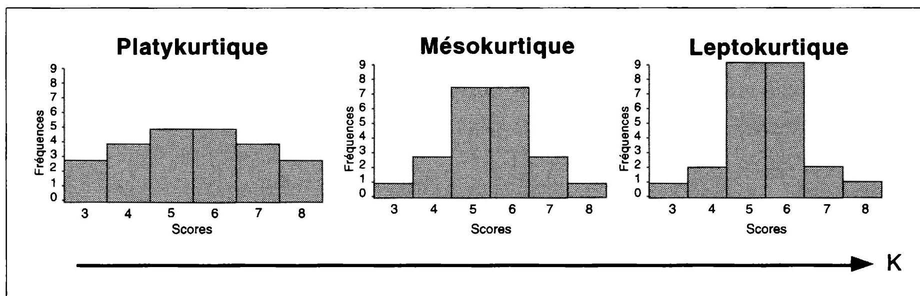


Figure 6 – Distributions en ordre croissant de kurtose

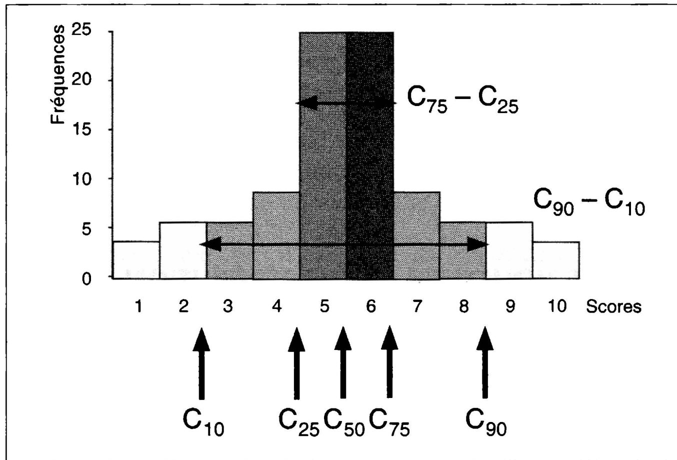


Figure 7 – Calcul de la kurtose à partir des rangs (quartiles et déciles) d'une distribution

La figure 7 nous indique que l'on peut avoir une idée du degré de voussure d'une distribution en calculant le rapport de deux étendues significatives. La première étendue porte sur l'intervalle semi-interquartile et nous renseigne sur le degré de dispersion des scores auprès de la moyenne. La seconde porte sur l'intervalle entre C_{90} et C_{10} et est davantage influencée par les valeurs extrêmes. Lorsqu'une distribution est leptokurtique, la première étendue est très petite par rapport à la seconde. Par contre, lorsque les valeurs sont fortement étalées, le rapport entre les deux étendues s'accroît. La formule suivante décrit un premier mode de calcul de la kurtose n'utilisant que les valeurs de position des scores. Cette formule est particulièrement adéquate dans le cas de mesures ordinales :

$$K = \frac{(C_{75} - C_{25})}{C_{90} - C_{10}} \tag{1.18}$$

Une distribution sera considérée comme mésokurtique lorsque la valeur de K sera voisine de 0,2632. Elle sera considérée comme leptokurtique lorsque $K < 0{,}2632$ et comme platykurtique lorsque $K > 0{,}2632$. Le principal avantage de cette formule est de nous permettre de nous faire rapidement une idée du degré de voussure d'une distribution à partir du calcul de quatre valeurs importantes de rangs centiles.

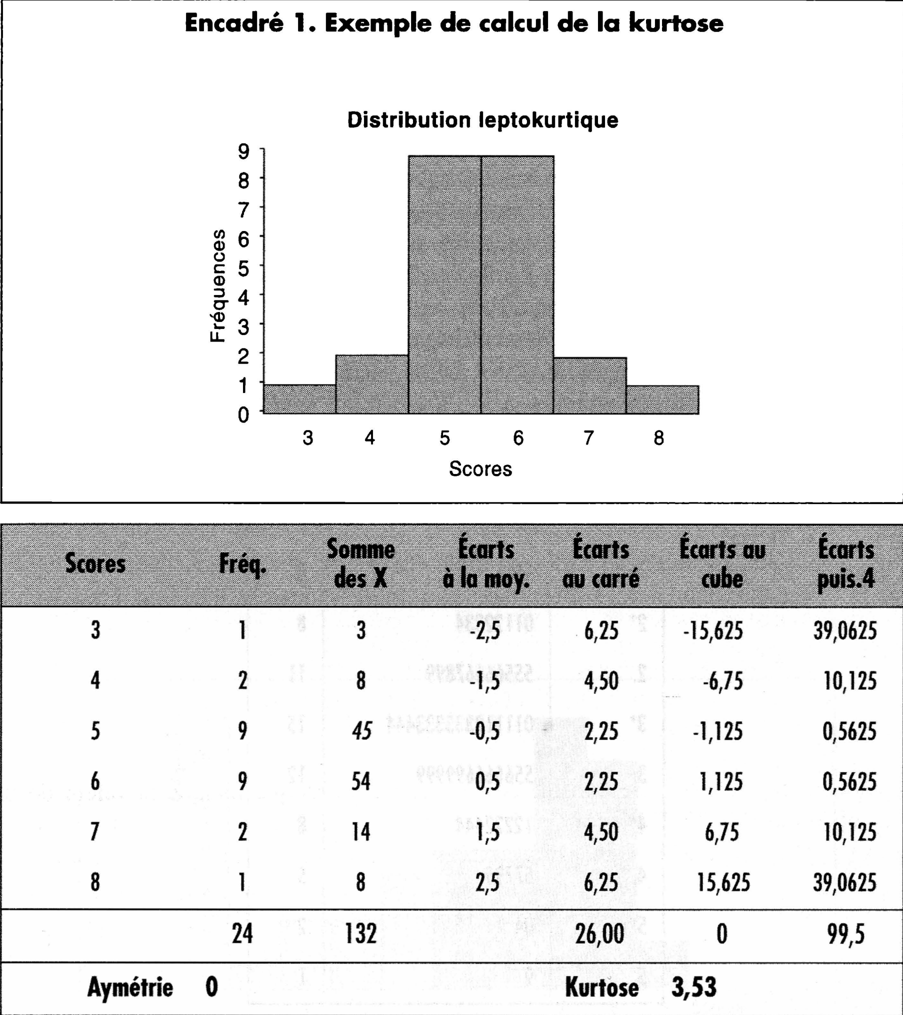
Pour obtenir une estimation plus précise de l'aplatissement d'une distribution, on peut, tout comme nous l'avons fait pour le calcul de l'asymétrie, mettre en rapport les différents moments d'une distribution. Le principal avantage de cette façon de procéder est de faire intervenir toutes les valeurs de la distribution et de tenir compte de leur importance relative. Le calcul de la kurtose en fonction des moments s'effectue de la manière suivante :

$$K = \frac{u_4}{u_2} = \frac{\sum x^4}{Ns^4} = \frac{N\sum x^4}{(\sum x^2)^2} \tag{1.19}$$

Dans l'équation (1.19), les termes ont la même signification que l'équation (1.17). Notez bien que la valeur de kurtose fournie par l'équation (1.19) n'est pas sur la

même échelle que celle de l'équation (1.18). Dans le cas de l'équation (1.19), une distribution est considérée comme mésokurtique lorsque $K = 3$. Lorsque $K > 3$, elle est leptokurtique et lorsque $K < 3$, elle est platykurtique. L'encadré 1 fournit un exemple de calcul de la kurtose à partir de l'équation (1.19).

Encadré 1. Exemple de calcul de la kurtose



Scores	Fréq.	Somme des X	Écarts à la moy.	Écarts au carré	Écarts au cube	Écarts puis.4
3	1	3	-2,5	6,25	-15,625	39,0625
4	2	8	-1,5	4,50	-6,75	10,125
5	9	*45*	-0,5	2,25	-1,125	0,5625
6	9	54	0,5	2,25	1,125	0,5625
7	2	14	1,5	4,50	6,75	10,125
8	1	8	2,5	6,25	15,625	39,0625
	24	132		26,00	0	99,5
Aymétrie	**0**			**Kurtose**	**3,53**	

2.6 REPRÉSENTATION GRAPHIQUE DES DONNÉES

Jusqu'ici, nous nous sommes restreints à un seul mode de présentation des données : l'histogramme des fréquences. Cette méthode de présentation graphique est adéquate dans la mesure où nous n'avons pas d'objection particulière à regrouper les données en catégories. L'histogramme de fréquences nous fournit alors un aperçu rapide de la distribution des résultats.

Lorsque nous voulons retenir les valeurs individuelles des données, le *diagramme en feuilles* constitue une alternative à l'histogramme des fréquences. Tout

comme l'histogramme, il repose sur un dénombrement des valeurs. Il existe plusieurs variantes de ce type de diagramme, mais pour l'essentiel il est constitué de *tiges* et de *feuilles*. Les tiges sont choisies pour regrouper les valeurs par tranches (de 10, de 100, etc.) sur lesquelles se greffent les feuilles en unités plus petites. La figure 8 représente un diagramme en feuilles typique.

Le diagramme en feuilles de la figure 8 se présente comme un histogramme que l'on aurait choisi de présenter horizontalement, couché sur son ordonnée. On reconnaît rapidement une distribution symétrique des résultats. Les tiges sont constituées des dizaines que l'on a séparées en deux : les dizaines associées aux valeurs 0 à 4 (1* 2* 3* 4* et 5*) et les dizaines associées aux valeurs 5 à 9 (1. 2. 3. 4. 5.). L'avantage de ce mode de présentation est de conserver les valeurs individuelles. C'est ainsi que nous réalisons qu'il n'y a ni valeur 46, ni valeur 48, ni valeurs de 51 à 53. Il est facile avec ce graphique de calculer toutes les valeurs de position. Sachant qu'il y a 69 valeurs, la médiane occupera donc le rang 35 [(69+1) / 2 = 35]. Comme les données sont déjà mises en ordre, il n'y a qu'à remonter d'une extrémité ou l'autre du diagramme jusqu'à la valeur dont le rang est 35 pour découvrir que la médiane est 33. Le même procédé permet de retrouver aussi rapidement les autres valeurs importantes de position telles que les quartiles, déciles ou autres.

Tige	Feuilles	Fr.
1*	00	2
1.	5566	5
2*	01122334	8
2.	55566667899	11
3*	011112333333444	15
3.	556666699999	12
4*	12222444	8
4.	57799	5
5*	04	2
5.	9	1

Figure 8 – Diagramme en feuilles

Parfois, par contre, nous ne sommes pas intéressés par l'ensemble des valeurs individuelles d'une distribution. C'est le cas lorsque la dispersion des valeurs constitue notre principale préoccupation, en particulier celle des valeurs extrêmes. Dans de tels cas, le *diagramme en boîte* constitue une alternative au diagramme en feuilles ou à l'histogramme de fréquences. Le diagramme en boîte illustre la dispersion des données autour de la médiane ainsi qu'aux extrémités. La boîte est définie à chaque extrémité par le premier et le troisième quartile (Q_3 *et* Q_1), et le trait à l'intérieur de la boîte représente la médiane. La boîte est prolongée à chaque extrémité par des traitillés ou *moustaches* au-delà desquelles se situent les valeurs extrêmes ou aberrantes.

La définition des valeurs extrêmes peut varier d'un auteur à l'autre. C'est pourquoi les diagrammes en boîte peuvent être différents selon les programmes de calcul. Une définition répandue, due à Tukey (1977), veut que l'on étende les moustaches jusqu'à la valeur la plus grande et jusqu'à la valeur la plus petite située à l'intérieur d'une étendue comprise entre la médiane et un écart égale à 1,5 fois l'intervalle interquartile (écart Q_3 - Q_1). Les valeurs à l'extérieur de cette étendue seront considérées comme des cas extrêmes et seront représentées par un symbole particulier, tel un astérisque (*).

La figure 9 présente le diagramme en boîte des données de la figure 5. Les extrémités de la boîte sont bien situées aux valeurs de $Q_1 = 4{,}5$ et de $Q_3 = 7{,}5$. Le trait intérieur représentant la médiane correspond bien à la valeur 6,5. L'étendue interquartile vaut 3 ($Q_3 - Q_1 = 7{,}5 - 4{,}5 = 3$). Les valeurs extrêmes seront donc situées au-delà de la valeur de la médiane plus ou moins un écart égale à 1,5 fois l'étendue de 3, soit 4,5. Les valeurs extrêmes seront soit inférieures à 2 (6,5 - 4,5 = 2) ou supérieures à 11 (6,5 + 4,5 = 11). Il n'y a qu'une seule valeur extrême, la valeur 1. Quant aux moustaches, elles s'étendent de la plus petite valeur à la plus grande valeur des données comprises entre 2 et 11 : soit entre 2 et 9, car il n'y a pas de valeur plus grande que 9.

Le diagramme en boîte de la figure 9 réussit bien à capter l'essence de la distribution des résultats. On voit clairement que la distribution est asymétrique négative et qu'elle comporte une valeur extrême. L'asymétrie est évidente au centre de la distribution, car la médiane ne se situe pas exactement au milieu de la boîte. Elle est également visible par l'étalement des moustaches, plus marqué vers les valeurs faibles.

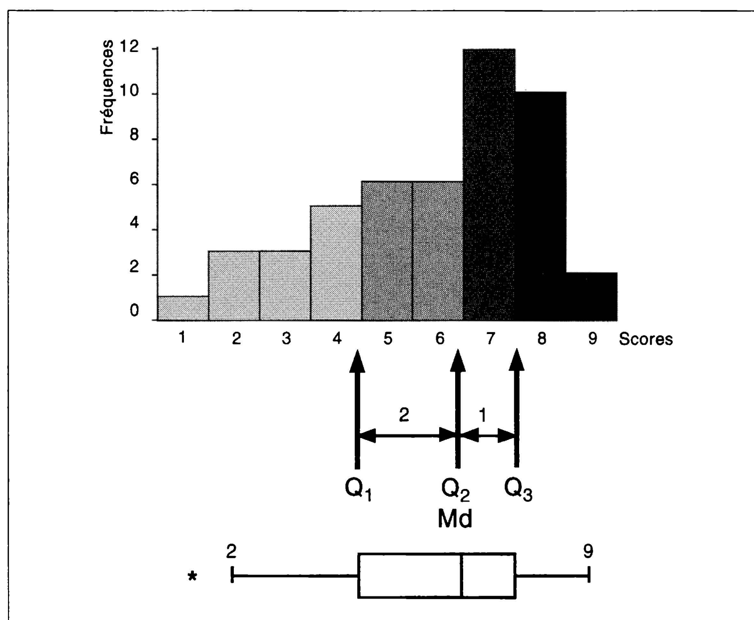


Figure 9 – Diagramme en boîte et histogramme de fréquences correspondant

Malgré ses avantages, le diagramme en boîte possède un inconvénient de taille : il ne permet pas de différencier les distributions possédant plus d'un mode. La représentation de l'étalement des valeurs autour de la médiane peut contribuer à voiler l'existence d'un second mode, comme c'est le cas d'une distribution bimodale. Seul un histogramme de fréquences ou un diagramme en feuilles pourrait nous révéler l'existence de plus d'un mode. Comme cette situation ne se produit que rarement, le diagramme en boîte demeure particulièrement attrayant par sa simplicité. Il faut toutefois être sensible à cette limite et au fait que des programmes de calcul différents peuvent définir autrement les valeurs extrêmes.

2.7 SYNTHÈSE ET APPLICATION

En résumé, pour tirer vraiment profit de l'étude d'une distribution de fréquences, nous avons besoin de calculer quatre valeurs qui nous permettent de la caractériser :

1. Une valeur de *tendance centrale* : c'est un indice de la valeur vers laquelle tend l'ensemble des résultats.
2. Une valeur de *dispersion* des résultats : c'est un indice du degré d'écart des résultats à la valeur de tendance centrale.
3. Une valeur de *symétrie* : cet indice permet de déterminer si les résultats se distribuent également de part et d'autre de la valeur de tendance centrale.
4. Un indice de *kurtose* : cet indice permet de déterminer si une proportion importante des résultats se regroupe autour de la valeur de tendance centrale ou si les résultats sont dispersés de manière plus ou moins égale dans l'ensemble de la distribution.

Pour calculer ces valeurs, il faut tenir compte de la nature de l'échelle de mesure des résultats puisque celle-ci limite les opérations et les transformations que l'on peut effectuer sur les nombres. Le tableau 3 présente un résumé de ces principaux indicateurs pour chaque échelle de mesure.

Tableau 3 – Les quatre caractéristiques d'une distribution selon l'échelle de mesure

	Tendance centrale	Dispersion	Symétrie	Kurtose
Échelle nominale	Mode			
Échelle ordinale	Médiane	Intervalle semi-interquartile	A (équation 1,16)	K (équation 1,18)
Échelle d'intervalle	Moyenne	Variance Écart-type	A (équation 1,17)	K (équation 1,19)

Il est important de déterminer si les caractéristiques d'une distribution de résultats correspondent bien à l'usage projeté. Dans bien des cas, une distribution normale, symétrique et mésokurtique, fera l'affaire. Elle représente en effet une situation

intermédiaire entre des cas extrêmes d'asymétrie et de voussure. Pourtant, il existe des situations où l'on préférerait obtenir un autre type de distribution afin de pouvoir mieux discriminer entre certains individus appartenant à une catégorie bien précise.

Dans les situations de sélection extrêmement compétitives, une distribution asymétrique négative est généralement préférable. Pour accorder un emploi par voie de concours, nous sommes intéressés par une distribution de fréquences où la plupart des participants auront des résultats très faibles et un très petit nombre de personnes des résultats élevés s'étalant sur la plus grande étendue possible de scores. C'est en donnant un test très difficile que l'on parvient généralement à obtenir une distribution asymétrique positive.

Parfois, comme dans les institutions scolaires, nous sommes intéressés à identifier le petit groupe d'élèves qui ne possèdent pas les pré-requis nécessaires d'apprentissage ou encore qui éprouvent des difficultés. Dans ce cas, nous aurons plutôt tendance à donner une examen très facile, qui sera réussi par la majorité des élèves et échoués par ceux-là mêmes qui éprouvent des difficultés. Un tel examen est fort susceptible de présenter une distribution de résultats asymétrique négative.

L'asymétrie nous permet d'accroître la discrimination à une seule extrémité d'une distribution. Dans le cas d'une évaluation-bilan, il peut être nécessaire de discriminer également aux deux extrémités d'une distribution. Par exemple, lorsqu'un psychologue utilise un test d'intelligence, il est intéressé d'obtenir le maximum de discrimination possible à chaque extrémité : autant parmi les valeurs très basses qui peuvent servir au classement en institution que parmi les valeurs très élevées qui peuvent décider d'une promotion ou d'un cheminement scolaire particulier. C'est dans ce genre de situation qu'il est préférable d'obtenir une distribution symétrique des résultats.

Le degré de voussure d'une distribution nous informe sur le degré de discrimination que l'on peut escompter sur l'ensemble d'une distribution et en particulier, au centre de celle-ci. Une distribution leptokurtique est le signe de résultats homogènes où il est très difficile de différencier les individus près de la moyenne. Une distribution platykurtique est le signe de résultats hétérogènes qui permettent de mieux différencier les individus au centre de la distribution. Par contre, la différenciation aux extrémités y est moins bonne.

Comme on peut le constater, toutes ces caractéristiques d'une distribution nous permettent de tirer des conclusions intéressantes sur la nature des résultats. Ces informations doivent être recoupées et leurs interactions étudiées de manière approfondie pour exploiter correctement toute l'information descriptive. L'application pratique qui suit vous permettra de juger de l'utilité de ces différents indicateurs.

Exemple

Tentons de voir comment il est possible de mettre à profit les informations concernant une distribution de scores dans le cas particulier de l'étude des résultats à un examen. Voici les résultats de 41 étudiants du baccalauréat inscrits à un cours optionnel de Docimologie d'une faculté d'éducation. Les scores ont été obtenus à l'examen de

mi-trimestre. Il y a dans ce groupe des étudiant(e)s (N=41) des programmes de sciences de l'éducation et du programme de sciences infirmières. Voici les résultats obtenus sur 20, dans un ordre quelconque :

9 17 16,5 13 9 16 15,5 4 10,5 15 15 18 9,5 12 16 13 15 14,5 13,5 15 16 17 15 12 12,5 13 14,5 13,5 14 14 14,5 14,5 14,5 6 7,5 8,5 9 10,5 10,5 10,5 13,5

Une telle série de nombres ne nous apprend que peu de choses. Tout au plus peut-on y noter la présence de deux valeurs très faibles (4 et 6) qui se détachent nettement du groupe. Mais comment en être sûr ?

Il nous faut examiner la distribution de scores. Étant donné qu'il y a 41 étudiants, que l'étendue entre le minimum (4) et le maximum (18) de cette distribution est de 15 (= 18,5-3,5), plusieurs regroupements de scores sont possibles. La figure 10 nous en propose trois : le premier en intervalles de classe de 1 crée trop de classes et ne comporte pas assez d'individus par classe. Par ailleurs, cette distribution n'est pas continue car il y a quelques classes dont la fréquence égale 0. Par contre, le regroupement en classes d'une étendue de 4 points est trop grossier. Le meilleur regroupement consiste en intervalles de classe d'une étendue de 2. Pour choisir le nombre de classes et l'étendue de chacune de celles-ci, nous tâchons de suivre les règles suivantes :

1. La distribution doit comporter un minimum de 7 classes et un maximum de 12.
2. La fréquence des résultats d'une classe ne devrait jamais être égale à 0.
3. Il faut limiter le nombre de classes dont la fréquence est inférieure à 5.

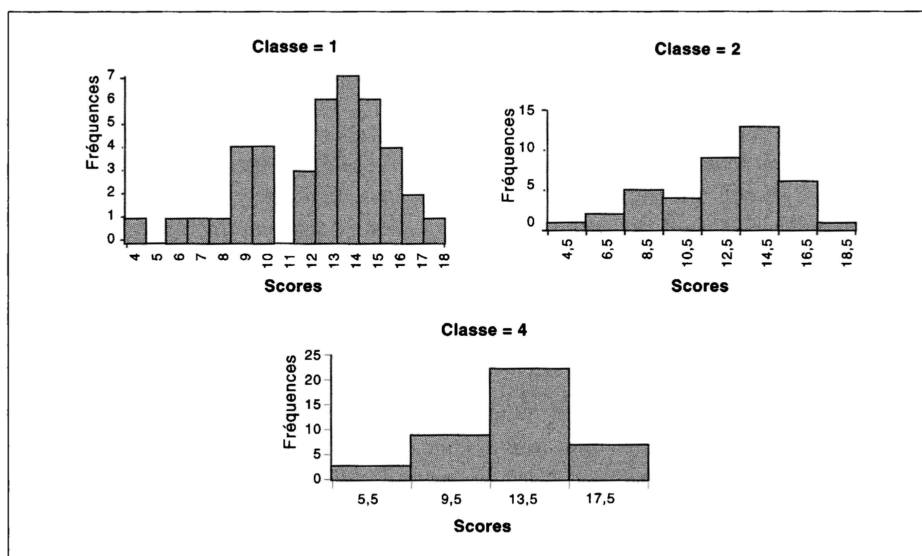


Figure 10 – Trois procédures de regroupement des scores en classes

Notre dernier choix satisfait ces exigences. De plus, il permet de constater que la distribution est asymétrique négative, justement à cause d'un petit nombre de valeurs très faibles. Ce genre de distribution convient-il bien à un examen mi-trimestre ? Fort

probablement oui, car il permet d'identifier les étudiants qui n'ont pas atteint les exigences du cours.

Ce dernier fait ressort plus clairement encore de l'observation de la figure 11 et plus particulièrement de la figure 12. Le diagramme en feuilles de la figure 11 nous permet de constater le caractère asymétrique de la distribution. Dans ce diagramme en feuilles, la présence de valeurs décimales (0,5) et la petite étendue de l'échelle de l'examen (20 points), ont compliqué le choix des tiges et des feuilles. La légende de cette figure indique que les tiges regroupent les feuilles par tranches de 2 et que chaque feuille représente une progression d'un demi point (0,5). Le résultat est fort similaire à l'histogramme de fréquences regroupant les données en classes de 2:

On remarque aussi que les deux résultats les plus faibles sont les valeurs 4 et 6. Ces deux valeurs sont bien des cas extrêmes, si l'on en juge d'après le diagramme en boîte de la figure 12. En effet, toutes deux se retrouvent en dessous de la moustache inférieure du diagramme : elles correspondent bien à des cas extrêmes. On voit aussi d'après ce diagramme en boîte que la distribution est asymétrique négative et que la médiane ne se situe pas exactement au centre de l'intervalle interquartile.

Tige	Feuilles	Fr
4	0	1
6	0$	2
8	#***$	5
10	###	3
12	00#***$$$	9
14	00#####*****$	13
16	000#**	6
18	0	6

Légende : 0 = + 0   * = +1   Exemples : 14 $ = 15,5   8* = 9
= + 0,5 $ = +1,5

Figure 11 – Diagramme en feuilles des données de l'exemple pratique

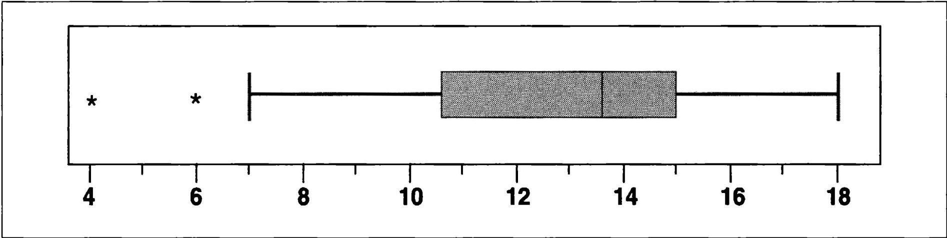


Figure 12 – Diagramme en boîte des données de l'exemple pratique

Le tableau 4 présente les principales statistiques descriptives calculées à partir des données. En consultant ces statistiques, on peut se demander si l'examen était de difficulté adéquate pour le groupe. La note de passage étant de 60% (12/20), une moyenne de 12,9/20 indique donc un examen difficile ou un groupe qui éprouve des difficultés. Il faut nuancer cette affirmation en tenant compte de l'asymétrie (-0,81 selon l'équation 1.18) des résultats : 50% des étudiants ont obtenus plus que 13,5 (la médiane).

On peut aussi tenter de déterminer si les résultats permettent de différencier les étudiants. Cinquante pour cent des résultats se situent entre 10,5 *(Q_1)* et 15 *(Q_3)* ce qui représente une étendue de 4,5. Une étendue de 4,5 peut sembler bien petite pour différencier 20 étudiants, mais il faut tenir compte que 4,5 représente près du quart de l'étendue totale de 20 points. D'après l'indice de kurtose de la distribution (équation 1.19), celle-ci serait de voussure moyenne (mésokurtique) ce qui indiquerait qu'il n'y a pas de concentration « *anormale* » des résultats autour de la moyenne.

À partir des résultats obtenus, on peut dire que l'examen a été légèrement difficile pour le groupe d'étudiants. L'asymétrie négative de la distribution a permis de faire ressortir clairement un petit groupe d'élèves très faibles ayant nettement échoué cet examen. Les résultats permettent aussi de différencier l'ensemble des élèves, en particulier près de la moyenne, même si c'est à cet endroit qu'il est le plus difficile de le faire. Dans le système universitaire canadien, l'examen se prête bien à la transformation des résultats en cotes A, B, C, D, E, car les scores obtenus couvrent pratiquement toute l'étendue de la distribution (minimum = 4 ; maximum = 18). Enfin, une distribution asymétrique négative révèle une légère accumulation de scores au-dessus de la moyenne. En règle générale, un professeur préférera obtenir une distribution asymétrique négative au lieu d'une distribution normale, surtout si l'objectif du cours n'est pas la sélection mais une approche fondée sur la pédagogie de la réussite.

Tableau 4 – Principales statistiques descriptives de l'exemple pratique

Moyenne	**12,88**
Médiane	**13,5**
Mode	**15**
Variance	**9,85**
Écart-type	**3,14**
Intervalle semi-interquartile	**2,25**
Asymétrie (équation 1.16)	**-1**
Asymétrie (équation 1.18)	**-0,81**
Kurtose (équation 1.19)	**3,1422**
Quartile 1	**10,5**
Quartile 3	**15**

3. La distribution normale

La distribution normale étant d'un usage très fréquent en psychométrie et en édumétrie, il est nécessaire de rappeler ses caractéristiques essentielles.

La distribution normale a été définie de manière précise par Laplace (1749-1827) et par Gauss (1777-1855). La première application de cette distribution à des données humaines (en l'occurrence la taille) a été réalisée par l'astronome belge Quetelet (1796-1874). La distribution normale est une distribution théorique d'une variable continue au sein d'une population infinie. Par conséquent, les distributions de fréquences que nous observons en psychologie et en éducation, basées sur un nombre fini de données discrètes, ne peuvent être qu'une approximation de cette distribution théorique.

Mathématiquement, la distribution normale est définie par la fonction suivante :

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2} \tag{1.20}$$

Dans cette équation, π et e sont des constantes ($\pi \cong 3{,}1416$ et $e \cong 2{,}7183$) ; μ et σ sont, respectivement, la moyenne et l'écart-type de la distribution dans la population. Si nous définissons une valeur pour μ et σ, nous pouvons alors calculer $f(X)$ pour toute valeur X. Les valeurs obtenues nous permettent de tracer la courbe normale théorique présentée dans la figure 13. Nous pouvons constater que la distribution normale est symétrique et unimodale. Ses limites sont $-\infty$ et $+\infty$. Par ailleurs, sa moyenne, son mode et sa médiane sont égales. Elles correspondent à la valeur se situant précisément au milieu de la distribution.

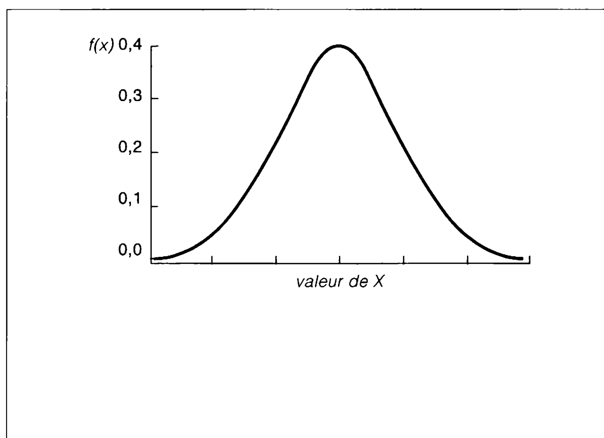


Figure 13 – La distribution normale théorique

Dans la mesure où X est une variable continue qui peut prendre une infinité de valeurs, il est impossible de calculer la probabilité d'occurrence d'une valeur précise de X. Par contre, nous pouvons évaluer la probabilité d'occurrence d'une valeur de X au sein d'un intervalle particulier. Cette probabilité correspond à l'aire sous la courbe

entre les deux bornes choisies. Elle peut être calculée par l'opération d'intégration de $f(X)$ entre les bornes x_i et x_j :

$$p\,(x_i \leq X \leq x_j) = \int_{x_j}^{x_i} f\,(x)\,d \qquad (1.21)$$

Heureusement, ce calcul fastidieux peut être évité en utilisant directement des tables de probabilité. Les tables existantes ont été élaborées en prenant 0 comme moyenne et 1 comme écart-type. Dans ce cas précis, la distribution normale est appelée *distribution normale réduite* (ou *distribution centrée réduite*) et les valeurs de X sont appelées scores z. Toute distribution normale, de moyenne et d'écart-type quelconques, peut être transformée en une distribution normale réduite au moyen de la formule suivante :

$$z = \frac{X - \bar{X}}{s_x} \qquad (1.22)$$

La transformation en *scores z* consiste simplement à calculer la différence entre chaque valeur de X et la moyenne de la distribution de X puis de diviser cette différence par l'écart-type de la distribution de X. Par exemple, si la moyenne de la distribution est 50 et son écart-type est 10, une valeur de X égale à 45 correspondra à une valeur z égale à -0,5. Soulignons que cette transformation est linéaire et qu'elle n'affecte pas les relations entre variables. Pour chaque valeur de X, nous avons seulement soustrait une constante et divisé par une constante. La forme de la distribution n'est pas modifiée par une telle transformation. Cela signifie que, si la distribution n'était pas normale avant transformation en scores z, elle ne le sera pas après. Contrairement à une idée répandue, cette transformation n'a pas la vertu de normaliser la distribution ! En fait, l'intérêt de cette transformation est de représenter toute distribution normale sur une échelle commune de moyenne égale à 0 et d'écart-type égal à 1. Il est ainsi possible d'utiliser la table de la distribution normale réduite quels que soient la moyenne et l'écart-type de la distribution normale originale.

Voyons à présent comment utiliser la table de probabilités de la distribution normale réduite. La table nous donne l'aire sous la courbe pour chaque intervalle entre la moyenne (c'est-à-dire 0) et les valeur de z qui s'échelonnent de 0,01 à 4,00. Par exemple (figure 14), pour l'intervalle entre la moyenne et 0,60, l'aire sous la courbe est égale à 0,2257. Cela signifie que, si nous tirons un score au hasard au sein de la distribution, nous avons un peu plus de 22% de chance de tirer un score inclus dans l'intervalle [0,00 ; 0,60].

Comme la distribution est symétrique, l'aire est identique pour les valeurs négatives de z. Si, par exemple (figure 14), nous voulons connaître la probabilité de tirer au sort un score compris entre -1 et -2, il nous suffit de regarder dans la table la valeur de l'aire pour les intervalles [0 ; 1] et [0 ; 2] puis de soustraire la première valeur de la seconde. Nous obtenons ainsi la probabilité de tirer au hasard un score situé entre 1 et 2, laquelle est identique à la probabilité de tirer au hasard un score situé entre -1 et -2. Concrètement, pour $z = 1$, l'aire sous la courbe est 0,3413 et pour $z = 2$, cette aire est 0,4772. Si nous soustrayons 0,3413 de 0,4772, nous obtenons l'aire pour

l'intervalle [1 ; 2] qui est égale à 0,1359. Cela signifie qu'aléatoirement, nous avons 13,59% de chance de tirer un score compris entre les valeur 1 et 2 de la distribution normale réduite. Ce pourcentage est le même pour l'intervalle [-1 ; -2].

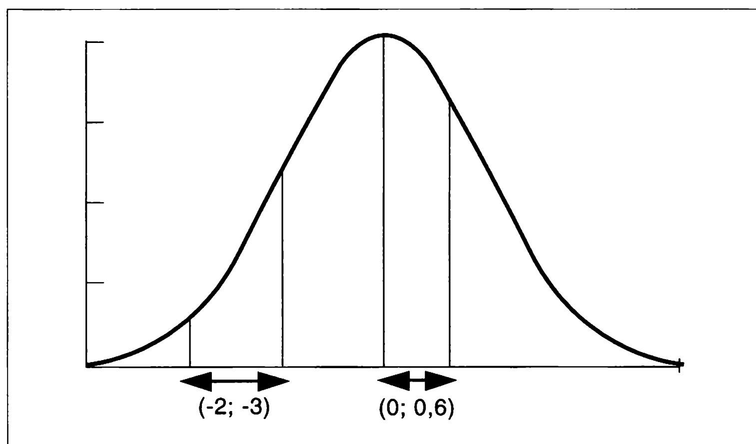


Figure 14 – Calcul de différentes aires sous la courbe

La table de distribution normale réduite nous permet de calculer des valeurs très utiles pour les praticiens. La figure 15 nous montre qu'au sein la distribution normale :

68,26% des scores sont inclus dans l'intervalle [-1σ ; +1σ],

95,44% des scores sont inclus dans l'intervalle [-2σ ; +2σ],

99,74% des scores sont inclus dans l'intervalle [-3σ ; +3σ].

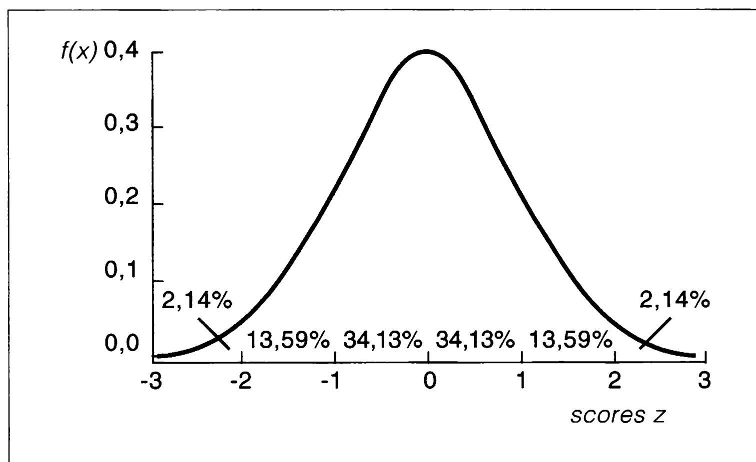


Figure 15 – Pourcentage de scores dans différents intervalles de la courbe normale

Nous ne devons pas perdre de vue que ces valeurs sont théoriques. Dans la pratique psychologique et éducative, nous ne mesurons que des variables discrètes et

nous n'obtenons qu'une approximation, souvent grossière, de la distribution normale théorique. La figure 16 illustre l'écart que l'on peut observer entre la distribution de données réelles et la distribution normale théorique.

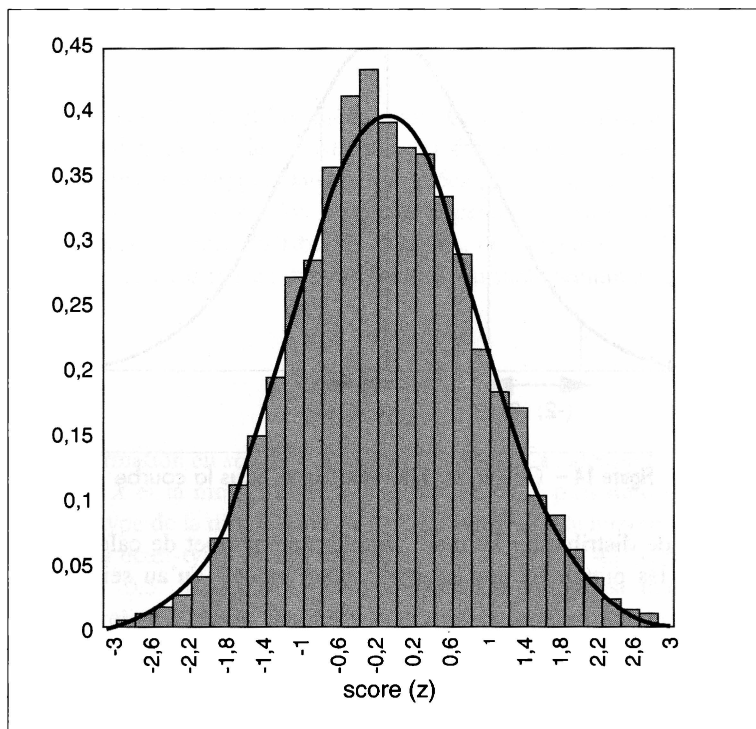


Figure 16 – Exemple d'écart entre la distribution réelle des scores et la distribution théorique

Plusieurs indicateurs peuvent nous informer de l'écart entre la distribution observée et la distribution théorique. Les deux plus utiles sont certainement les indices d'asymétrie et de voussure qui sont présentés de manière détaillée dans la section 2 de ce chapitre.

4. Conclusion

Ce premier chapitre n'a certes pas fait le tour de toutes les méthodes de description des données. Elles sont fort nombreuses et correspondent à des besoins particuliers. Celles qui ont été présentées sont les plus courantes : elles permettent de décrire les données dans différentes situations en tenant compte des objectifs visés et de la nature des échelles de mesure.

Chaque méthode de présentation graphique des données, de calcul des valeurs de tendance centrale, de dispersion, de symétrie, de voussure possède ses propres avantages ainsi que ses inconvénients. C'est pourquoi plusieurs méthodes doivent être employées en conjonction les unes avec les autres afin d'offrir une perspective

d'ensemble qui soit valide et fiable. Il est également important de reconnaître l'algorithme sur lequel se base le calcul des différentes caractéristiques d'une distribution. Plusieurs programmes de calcul, utilisant des algorithmes différents, peuvent présenter des valeurs calculées de symétrie ou de kurtose différentes. Dans de tels cas, il faut consulter la documentation fournie avec le logiciel pour en être position d'interpréter correctement les résultats.

CHAPITRE 2

NOTIONS D'INFÉRENCE STATISTIQUE

Les méthodes présentées jusqu'à présent permettent de décrire un échantillon ou encore toute une population à condition que nous puissions avoir accès à tous ses membres. Ce n'est pas toujours possible. En éducation et en psychologie, nous avons souvent pour objectif de connaître une population à partir d'un échantillon représentatif de ses membres. C'est là le domaine des *statistiques inférentielles* qui feront l'objet du présent chapitre.

1. Échantillon et population

La mesure, qu'elle soit *critériée* ou *normative*, repose généralement sur des estimations. En effet, on ne peut questionner un individu particulier sur tous les items d'addition, pas plus que l'on ne peut comparer la réussite de tous les individus pour lesquels un test d'addition a été développé. Nos conclusions s'appuient généralement sur les estimations que nous faisons au moyen :

1. d'un échantillonnage d'items selon des critères précis, dans le cas de la *mesure critériée* ;
2. d'un échantillonnage représentatif de personnes, dans le cas de la *mesure normative*.

Chaque type de mesure accorde donc priorité à un type d'*échantillonnage* : échantillonnage des items de l'univers de contenu, en mesure critériée ; échantillonnage des personnes de la population d'intérêt, en mesure normative. Traditionnellement, l'éducation s'est particulièrement intéressée au premier problème d'échantillonnage. La psychométrie, pour sa part, s'est surtout attachée au deuxième. Ceci se traduit par des procédures différentes de construction des tests.

En éducation, ou plus précisément en édumétrie, la définition a priori de l'univers de contenu à mesurer a pour effet que le principal travail de sélection des items se fait avant le testing. En psychométrie, lorsque les résultats des tests sont employés pour différencier des individus entre eux, il est parfois très difficile de savoir à l'avance quels items vont accroître la discrimination entre les personnes. Ce n'est qu'a posteriori qu'une sélection des items peut véritablement avoir lieu, soit une fois que ceux-ci ont été administrés à un premier échantillon représentatif de la population d'intérêt.

Dans la pratique, si nous souhaitons différencier des individus en fonction de leur intelligence, nous chercherons à utiliser des items qui nous permettent de discriminer dans toute la population. Il ne serait pas approprié de mettre à l'essai notre test sur un échantillon restreint de la population comme, par exemple, les étudiant(e)s de niveau universitaire ou les élèves de classes spéciales, à moins que notre but ne soit précisément d'établir des différences parmi les individus de chacune de ces sous-populations. Si nous voulons discriminer dans l'ensemble de la population, nous chercherons plutôt à obtenir un échantillon *représentatif* de toute la population. Pour ce faire, il existe plusieurs méthodes d'échantillonnage plus ou moins bien adaptées à différents problèmes d'estimation. Celles-ci seront décrites en détail dans le chapitre 7.

En psychologie différentielle et en psychométrie, la comparaison entre individus est très importante. Nous allons délaisser temporairement le problème de la différenciation des processus cognitifs, qui sera abordée dans le chapitre 6, pour nous concentrer sur celui de la différenciation des personnes. Historiquement, cette problématique est plus ancienne et pose le problème de l'estimation d'une norme à laquelle sont comparés tous les individus d'une même population. Cette norme est généralement la moyenne de la population des individus. L'estimation de cette moyenne au moyen d'un échantillon représentatif est donc de première importance, car cette norme est la valeur par rapport à laquelle chaque personne sera comparée.

Parce que les valeurs de l'échantillon et de la population correspondent à des réalités différentes, les conventions en statistiques veulent que les paramètres d'une population soient exprimés au moyen d'une lettre grecque, alors que les paramètres de l'échantillon sont exprimés par la lettre correspondante de l'alphabet romain. La moyenne de la population s'écrit donc μ et la moyenne de l'échantillon s'écrit m. L'écart-type de la population s'écrit σ, alors que celui de l'échantillon s'écrit s.

Lorsque nous décrivons des valeurs estimées, les conventions veulent que nous utilisions une lettre grecque accompagnée d'un accent circonflexe. Par exemple, on écrira $\hat{\sigma}_X^2$ pour signifier la variance de la population estimée à partir de la variance de l'échantillon des valeurs de X. Toutefois, pour alléger la notation algébrique, nous décrirons de la même façon, valeurs de l'échantillon et valeurs estimées à partir de l'échantillon, au moyen des caractères romains. Ainsi, s_X^2 signifiera tout autant, *variance de l'échantillon* que *variance de la population calculée à partir de l'échantillon*. Le contexte sera habituellement suffisant pour distinguer ces deux situations lorsque ce sera nécessaire.

1.1 INFÉRENCE ET ESTIMATION

L'inférence fait partie des opérations mentales à notre disposition pour saisir une information non présente. Legendre (1993) définit l'inférence comme un "mode de raisonnement qui consiste à tirer une conséquence ou une conclusion logique d'un ensemble de données". Ce mode de raisonnement est relativement fréquent et plutôt familier dans les situations de la vie courante, "mais dans les cas où certains domaines du savoir s'éloignent des lieux communs et présentent un degré d'abstraction élevé, ou si ces domaines ne sont pas suffisamment familiers au sujet, il lui devient particulièrement difficile de faire les inférences demandées" (Legendre, 1993, p.714). C'est le cas notamment en statistiques.

Pour mieux saisir cette notion d'inférence, faisons appel à une situation de la vie quotidienne. Supposons que vous vous promenez dans votre quartier. Vous ne vous attendez pas à croiser sur votre chemin une personne mesurant plus de 2 mètres. Si, avant votre départ, on vous demandait de faire une prédiction à propos d'un tel événement, vous parieriez probablement que vous ne rencontrerez pas une telle personne et vous auriez une grande confiance en votre prédiction.

Votre assurance repose sur une inférence très simple. Vous connaissez bien les gens qui habitent votre quartier et vos observations antérieures lors de vos nombreuses promenades vous ont appris qu'il n'y a personne de cette taille dans votre environnement. Pour que vous croisiez une personne mesurant plus de 2 mètres, cette personne devrait provenir de l'extérieur du quartier et se promener au même moment que vous. Vous en concluez que la probabilité de rencontrer une personne mesurant plus de 2 mètres est tellement faible que vous préférez rejeter cette possibilité.

En inférence statistique, nous raisonnons de la même manière. Nous estimons les probabilités qu'un événement se produise au hasard avant de prendre une décision. Si un événement a très peu de chances de se produire au hasard, alors nous préférons accepter une autre hypothèse, *l'hypothèse alternative*, selon laquelle l'événement dont nous sommes témoins est imputable à autre chose que les simples fluctuations aléatoires. Toutefois, aucune des décisions que nous prenons dans le contexte de l'inférence statistique n'est absolument certaine, puisque nous fondons notre prise de décision sur des probabilités. Il y a donc un risque d'erreur associé à chaque décision et les tests statistiques nous permettent de l'estimer.

1.2 ÉCHANTILLONNAGE ET ESTIMATION DE LA MOYENNE D'UNE POPULATION

En statistiques, nous ne sommes pas seulement intéressés par le calcul des paramètres d'un échantillon. En effet, les sondages électoraux seraient bien peu intéressants si tout ce qu'ils nous apprenaient ne se limitait qu'aux intentions de vote des seules personnes sondées. Il en va de même de nombreuses caractéristiques humaines qui sont mesurées en éducation et en psychologie. Bien souvent les caractéristiques de l'échantillon ne nous intéressent que dans la mesure où elles sont représentatives de la population entière dont est tiré l'échantillon.

Pour qu'un échantillon soit représentatif de la population, les membres de la population doivent être choisis au hasard avec une chance égale d'être sélectionnés. Nous nous limiterons ici à la méthode d'échantillonnage aléatoire simple. Cette méthode nous permet d'obtenir un échantillon représentatif de la population. Ceci ne signifie pas que les caractéristiques de l'échantillon soient exactement celles de la population. L'échantillon permet seulement d'estimer les caractéristiques de la population avec une marge d'erreur plus ou moins grande. Plus nous sélectionnons une proportion importante de la population, plus nous pouvons avoir confiance dans cette estimation.

Par exemple, pour déterminer la qualité de l'eau d'un lac, il ne suffira pas de puiser l'eau à un seul endroit. Il faudra prendre des échantillons d'eau à différents points du lac et à des profondeurs différentes. Pour ne pas biaiser notre échantillon, nous choisirons ces endroits et ces profondeurs au hasard. Plus nous puisons l'eau à des endroits variés choisis au hasard, plus nous pouvons avoir confiance en notre estimation de la qualité de l'eau. Il en va de même lorsque nous tentons, par des techniques d'échantillonnage, d'estimer les caractéristiques d'une population entière. Par exemple, nous pouvons nous demander quel est le score moyen d'indépendance du champ (*field independence*) d'élèves de cinquième année. Au lieu d'interroger tous les élèves de cinquième année — ce qui pourrait s'avérer irréaliste ou impossible pour toutes sortes de raisons pratiques et économiques — nous choisissons de ne retenir qu'un échantillon représentatif de ceux-ci, tiré au hasard de la population.

Quelle serait la moyenne, la variance de la caractéristique "indépendance du champ" estimée sur base de notre échantillon ? Les lois de l'inférence statistique nous apprennent que la meilleure estimation de la moyenne de la population est la moyenne de notre échantillon. Nous exprimerons ce premier principe par l'équation suivante, où m représente la moyenne de l'échantillon et μ celle de la population :

$$m \cong \mu \qquad (2.1)$$

Toutefois, nous n'avons aucune certitude que la moyenne m de notre échantillon soit véritablement celle de la population. Si notre échantillon a été tiré au hasard, il est possible d'évaluer la probabilité que la moyenne de l'échantillon soit différente de celle de la population. Sur cette base, nous pouvons construire un *intervalle de confiance* autour de la moyenne de l'échantillon à l'intérieur duquel la moyenne de la population a une certaine probabilité de se trouver. Pour déterminer cet intervalle de confiance, il nous faut connaître la variance des moyennes des échantillons tirés au hasard au sein de la population. Or, le bon sens nous incite à croire que plus les échantillons tirés de la population seront grands, plus petite sera l'incertitude entourant l'estimation de la moyenne de la population. De fait, les lois de l'inférence statistique nous indiquent que la variance des moyennes s_X^2 calculée à partir d'échantillons aléatoires de taille n, sera n fois plus petite que la variance $s_{\bar{X}}^2$ des n scores tirés de l'échantillon. L'équation suivante représente ce deuxième principe :

$$s_{\bar{X}}^2 = \frac{s_X^2}{n} \qquad (2.2)$$

L'estimation de la variance des moyennes d'échantillons de taille *n* constitue ce que l'on appelle l'*erreur d'estimation de la moyenne*. Puisque les moyennes des échantillons se distribuent normalement, il nous est donc possible de calculer un intervalle de confiance autour de la moyenne de l'échantillon à l'intérieur duquel existe une probabilité de 95% de retrouver la moyenne de la population.

Appliquons le calcul de l'erreur d'estimation de la moyenne au problème de l'estimation du quotient intellectuel moyen d'un groupe de 100 élèves tirés au hasard. Nous savons que les quotients d'intelligence se distribuent dans la population avec une moyenne de 100 et un écart-type de 15 (c'est le cas des Q.I. calculés au moyen de l'échelle Weschler). L'erreur-type de la moyenne est obtenue au moyen du calcul suivant :

$$s_{\bar{X}}^2 = \frac{s_X^2}{n} = \frac{225}{100} = 2{,}25$$
$$s_{\bar{X}} = \sqrt{2,22} = 1{,}5 \qquad (2.3)$$

L'erreur d'estimation nous permet de reconstruire la distribution des moyennes d'échantillons de 100 sujets tirés d'une population de moyenne 100 et d'écart-type 15. Cette distribution des moyennes aura pour moyenne globale la même valeur, 100, et pour écart-type l'erreur d'estimation 1,5. Selon les probabilités de la loi normale et *lorsque la taille des échantillons est supérieure à 30*, il y a 95% de chances que la moyenne de l'échantillon se trouve dans un intervalle compris entre ±1,96 $s_{\bar{X}}$ ce qui, dans l'exemple, est égal à (± 1,96 x 1,5) = ± 2,94.

En conclusion, un chercheur qui prétendrait tirer un échantillon représentatif de la population du point de vue du quotient d'intelligence et qui obtiendrait à partir d'un groupe de 100 sujets sélectionnés au hasard un quotient intellectuel moyen de 105, pourrait difficilement prétendre que son échantillon a été tiré de la population décrite précédemment puisqu'elle se situe en dehors de l'intervalle de confiance de 95% compris entre 100 ± 2,94 (entre 97,06 et 102,94). Un groupe de 100 sujets dont la moyenne des QI serait de 105 a donc moins de 5 chances sur 100 d'être tiré au hasard. Cet événement statistique est possible, mais il est très rare. C'est pourquoi le chercheur tirera la conclusion que l'échantillon n'a pas été tiré au hasard et qu'il n'est pas représentatif de la population. En prenant cette décision, le chercheur risque de se tromper. En effet, chaque fois que de tels échantillons - bien que rares - sont tirés au hasard, le chercheur se trompera. Toutefois, ce risque d'erreur est inférieur à 5%. Nous verrons dans la section suivante comment le risque d'erreur influence notre prise de décision et la puissance de nos tests statistiques.

1.3 INFÉRENCE STATISTIQUE ET LOIS DE PROBABILITÉ

La figure 1 illustre la différence entre *erreur d'estimation* et *écart-type*. Lorsque les moyennes des Q.I. sont calculées à partir de grands échantillons (dans ce cas-ci n=100) l'erreur d'estimation de la moyenne est beaucoup plus petite que l'écart-type des scores brutes. Même si un Q.I. de 105 est relativement fréquent dans une distribution de scores brutes, un Q.I. moyen de 105, calculé sur une centaine de sujets tirés au hasard, a une probabilité très faible.

Les calculs précédents valent pour des échantillons de grande taille (n supérieur à 30). Dans de tels cas, la *loi normale* sert au calcul des probabilités des moyennes. Lorsque les échantillons comportent moins de 30 sujets, l'estimation de la variance de la population a plus de chances d'être sous-estimée que surestimée. Les valeurs extrêmes de la population contribuant de façon importante à la variance de la population risquent peu de se retrouver dans un petit échantillon. Pour de petits échantillons, la distribution des moyennes ne suit pas exactement la loi normale, mais une distribution platykurtique, *la loi t de Student.* L'intervalle de confiance à 95% est alors supérieur à $\pm 1{,}96\ \sigma_{\bar{X}}$ et correspond à $\pm\ t_{0{,}05}\ \sigma_{\bar{X}}$. La valeur de t est obtenue en consultant les tables de probabilités de *t de Student* pour le nombre de sujets de l'échantillon moins 1 (soit le nombre de *degrés de liberté* que nous expliquerons plus loin).

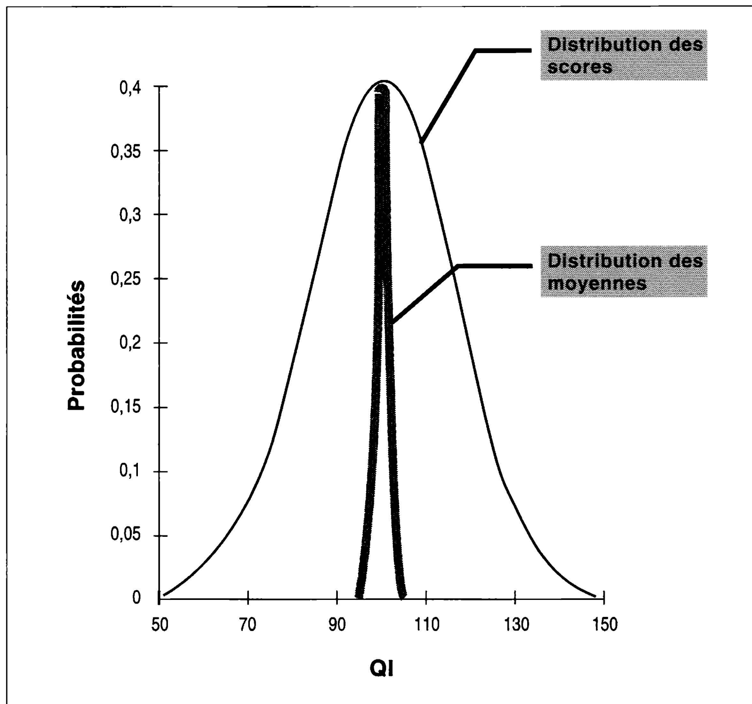


Figure 1 – Erreur d'estimation et écart type des Q.I. pour n=100

Si nous appliquons les données de l'exemple précédent au cas d'un échantillon de 17 sujets tirés de la même population, l'erreur d'estimation de la moyenne sera :

$$s_{\bar{X}}^2 = \frac{s_X^2}{n} = \frac{256}{16} = 16 \qquad (2.4)$$
$$s_{\bar{X}} = \sqrt{16} = 4$$

Comme prévu, l'erreur d'estimation de la moyenne est beaucoup plus grande à cause de la taille réduite de l'échantillon. De plus, du fait de l'incertitude plus grande entourant l'estimation de la variance de la population, l'intervalle de confiance sera

supérieur à l'intervalle habituel de ±1,96 pour de grands échantillons. Nous devons calculer un nouvel intervalle à partir de la valeur de $t_{0,05} > 1,96$ pour un nombre de degrés de liberté *(dl)* égal à *17-1 =16*. Le nouvel intervalle calculé sera égal à ± 2,131 $s_{\bar{X}}$, soit ± 8,524.

Comme on peut le constater, la marge d'incertitude s'est beaucoup accrue en utilisant un échantillon plus petit. Avec un échantillon de 100 sujets, nous réduisions considérablement la possibilité qu'une moyenne de 105 puisse provenir d'une population dont la moyenne est égale à 100. Avec un échantillon de 30 sujets maintenant, la même valeur (105) se situe à l'intérieur de l'intervalle de confiance à 95%. Nous serions donc prêts à accepter qu'une moyenne de 105 puisse provenir d'une population dont la moyenne est 100.

1.4 Inférence statistique et prise de décision

L'inférence statistique nous permet de faire davantage que le simple calcul de l'erreur d'estimation de la moyenne de la population. Nous pouvons tenter de déterminer si les moyennes de deux échantillons peuvent être considérées comme différentes. Cette question revient à se demander si la différence observée entre deux moyennes est probable en partant du postulat que les deux échantillons ayant servi au calcul des moyennes ont été tirés de la même population. S'il est peu probable que les moyennes des deux échantillons aient été tirées de la même population, alors nous considérons qu'un facteur quelconque est intervenu pour créer cet écart entre les deux moyennes — en d'autres termes, pour *biaiser* l'estimation de l'une des deux moyennes.

Un exemple permettra de mieux comprendre la situation précédente. Supposons qu'en vous promenant dans la rue vous faites la rencontre de deux personnes, l'une mesurant 1,9 m et l'autre 1,7 m. Rien de surprenant là-dedans puisque 1,9 m et 1,7 m sont des hauteurs probables dans la population normale. Supposons, toutefois, que vous rencontriez 20 personnes dont la hauteur moyenne est de 1,9 m, puis 20 autres dont la hauteur moyenne est de 1,7 m. Vous commencez à vous interroger. Si les valeurs individuelles de 1,7 m et 1,9 m ont des chances raisonnables de se produire dans la population, un écart de 20 cm entre deux groupes de 20 personnes l'est très peu. Tout vous portera à croire que les individus de chacun de ces groupes ne sont pas représentatifs de la population en général et que cette différence de 20 cm, pourtant normale entre deux individus, ne l'est pas entre deux groupes. Ce serait d'autant plus vrai si ces deux groupes sont formés d'un grand nombre d'individus tirés au hasard. Dans ce cas-ci, vous pourriez avoir assisté à la sortie des membres d'une équipe de basket-ball, suivie quelques minutes plus tard par celle d'un groupe de karaté. Ces groupes sont différents et ne peuvent donc être considérés comme tirés de la même population.

2. Comparaison de deux moyennes

Deux techniques statistiques apparentées à la loi *t de Student* nous permettent de calculer la probabilité des différences entre deux moyennes. Elles permettent toutes de répondre à la question suivante : *à partir de quel moment peut-on considérer deux moyennes comme significativement différentes l'une de l'autre ?* Pour répondre à cette

question, il faut connaître la probabilité que de telles différences entre moyennes se produisent au hasard lorsque les deux moyennes proviennent de la même population ou de deux populations dont la moyenne est identique. La loi de probabilité du t de Student — que nous venons d'étudier dans le cas de l'estimation de la moyenne pour de petits échantillons — permet le calcul des probabilités de ces différences entre moyennes. Il existe deux façons de calculer la valeur de t pour la comparaison de deux moyennes :

1. la méthode pour deux *échantillons indépendants* ;
2. la méthode pour deux *échantillons pairés*, dite aussi *des échantillons liés.*

La première méthode est la plus simple. Nous tirons au hasard deux échantillons, indépendamment l'un de l'autre, dont nous calculons les moyennes. Il s'agit alors de calculer l'écart entre les deux moyennes. La seconde méthode introduit un élément supplémentaire. Plutôt que de comparer les deux groupes dans leur ensemble, il s'agit de comparer les individus des deux groupes par paires, en choisissant de calculer la différence entre les résultats obtenus entre les individus d'une même paire, puis de calculer la moyenne de ces différences. Bref, la méthode pour échantillons indépendants vise à déterminer si *la différence entre les moyennes* de deux groupes est significative, alors que la méthode pour échantillons pairés vise à déterminer si *la moyenne des différences* est significative.

Pour que la méthode pour échantillons pairés ait un sens et qu'elle donne lieu à des résultats réellement différents de la méthode pour échantillons indépendants, il faut que le *pairage* entre les sujets soit pertinent. C'est le cas lorsque, pour déterminer la valeur de deux méthodes d'apprentissage, nous comparons les résultats d'individus de mêmes quotients d'intelligence. Nous savons que les capacités d'apprentissage sont fortement influencées par les aptitudes intellectuelles. En ne comparant que les résultats d'individus de mêmes aptitudes, nous éliminons la possibilité que les différences observées entre les résultats soient imputables à cette variable. La comparaison que nous faisons alors entre les deux groupes est d'autant plus pertinente. Par contre, si nous avions choisi de pairer les individus selon leur taille, il est fort probable que la comparaison n'aurait rien apporté puisque la taille n'a aucune influence sur l'apprentissage.

Le pairage des sujets vaut également lors de mesures répétées. Le sujet est alors comparé à lui-même. Cette situation se rencontre lorsque nous souhaitons étudier le progrès individuel en éducation. C'est le cas aussi des protocoles expérimentaux de type pré-post traitement que ce soit en psychologie ou en sciences de l'éducation.

Le pairage des sujets permet d'effectuer de meilleures comparaisons, en particulier lorsque les échantillons sont petits. Plus les échantillons sont petits, plus il est possible de rencontrer accidentellement deux groupes dont les aptitudes intellectuelles sont différentes. Or, cette seule différence dans les aptitudes intellectuelles peut expliquer, totalement ou en partie, l'écart dans les résultats d'apprentissage des deux groupes. Le pairage permet d'éliminer cette possibilité, au prix cependant, d'un travail plus complexe d'échantillonnage. Tout comme la méthode pour échantillons indépendants, les sujets seront tirés au hasard. Puis, des paires de sujets semblables — à l'intérieur d'une certaine marge de tolérance — seront constituées. Par exemple, on considérera comme de même niveau d'intelligence deux personnes dont le QI se situe entre 105 et

110. Le hasard interviendra à nouveau pour déterminer à quel groupe sera assigné chaque membre de la paire. Le pairage des sujets peut donner lieu à des difficultés imprévues. Pour constituer des paires de sujets comparables, il peut être nécessaire de tirer plusieurs sujets. Mais cet effort en vaut la peine. Dans la mesure où la variable de pairage exerce une influence réelle sur les données des deux groupes, la comparaison entre les deux groupes s'en trouve améliorée. En termes statistiques, nous dirons que la méthode pour deux échantillons pairés, lorsqu'elle s'avère pertinente, donne lieu à un test plus *puissant* des différences entre les deux groupes.

Le tableau 1 présente un exemple employant les deux méthodes. Dans le cas d'échantillons indépendants, il n'est pas possible d'identifier entre quels sujets les écarts entre les deux groupes peuvent être calculés. C'est pourquoi, la moyenne des deux groupes est calculée sur l'ensemble des sujets et la différence est établie entre les deux moyennes. Dans le cas d'échantillons pairés, l'écart est calculé pour chaque paire et c'est la moyenne des écarts qui sert d'indicateur de la différence entre les deux groupes. Dans l'exemple du tableau 1, les mêmes données ont été employées dans chaque groupe.

Tableau 1 – Comparaison de deux moyennes.
Méthodes pour échantillons indépendants et pairés

Échantillons indépendants			**Échantillons pairés**			
	Groupe 1	*Groupe 2*	*Paires*	*Groupe 1*	*Groupe 2*	*Dfférences*
	10	14	1	10	15	-5
	12	15	2	12	14	-2
	21	19	3	13	15	-2
	17	19	4	16	17	-1
	16	18	5	16	19	-3
	16	17	6	17	18	-1
	13	15	7	18	19	-1
	18	20	8	21	20	1
moyennes	15,38	17,13	moyennes	15,38	17,13	-1,75
écarts-types	3,54	2,23	écarts-types	3,54	2,23	1,75
erreurs d'estimation	1,25	0,79	erreurs d'estimation	1,25	0,79	
valeur de *t*		0,95	valeur de *t*			2,29
dl		14,00	dl			7,00
probabilité		0,26	probabilité			0,03

Le tableau 1 présente également les valeurs de *t* pour chaque méthode. La valeur *t* est une mesure de la différence entre les deux groupes qui tient compte de leurs moyennes et de leurs variances respectives. Moins les distributions des deux

groupes se chevauchent, plus leurs moyennes sont séparées l'une de l'autre, plus la valeur de t est élevée, quelque soit la méthode par laquelle elle est calculée. Lorsque la valeur de t est élevée, il y a peu de chance que les moyennes des deux groupes proviennent de la même population. C'est ce qu'indiquent les probabilités associées à chacune des valeurs de t calculées dans le tableau 1.

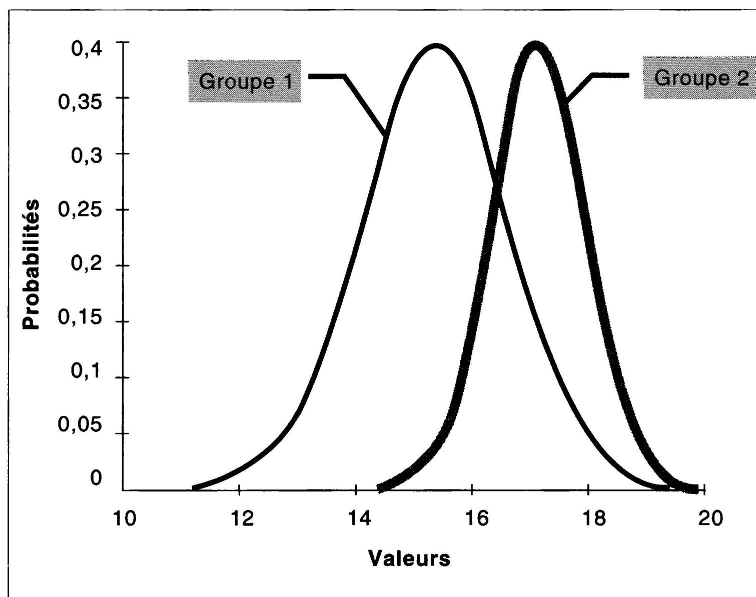


Figure 2 – Distribution des moyennes de deux échantillons indépendants (données du tableau 1)

La figure 2 présente les distributions normales des moyennes de chaque groupe en tenant compte de leurs erreurs d'estimation respectives. Comme on peut le constater, il y a peu de chevauchement entre les deux distributions de moyennes. Il y a donc peu de chance qu'elles proviennent toutes deux de la même population. Ce graphique illustre également qu'il y a deux façons de réduire le chevauchement entre les deux distributions. La plus simple, sans aucun doute, est d'accroître l'écart entre les moyennes des deux groupes. La seconde, moins évidente, est de réduire l'erreur d'estimation, dont la variance est N fois plus petite que celle de l'échantillon. En choisissant des échantillons plus grands, l'erreur d'estimation aurait été plus petite et le chevauchement encore moindre.

La valeur de t pour échantillons indépendants se calcule au moyen de l'équation suivante :

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.5}$$

où le numérateur indique la différence entre les moyennes des deux groupes $(\bar{X}_1 - \bar{X}_2)$ et où s_1^2 et s_2^2 représentent les variances de chaque échantillon indépendant de taille n_1 et n_2.

La valeur de t pour échantillons pairés se calcule différemment. Elle fait intervenir une nouvelle valeur, D, qui est l'écart entre les deux valeurs de chaque paire. Dans l'équation (2.6), $\bar{D}$ représente la moyenne des différences de chaque paire, s_D l'écart-type des valeurs de différences et n représente le nombre de paires.

$$t = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}} \tag{2.6}$$

Quelle que soit la manière de calculer la différence entre deux échantillons, t est le résultat d'une transformation mathématique de la différence qui nous permet d'en estimer la probabilité. Pour connaître cette probabilité, il faut aussi tenir compte de la valeur des *degrés de liberté (dl)*. Cette valeur *dl* indique le nombre de résultats libres de varier dans chaque situation. Elle se retrouve dans tous les tests d'inférence statistique et est absolument nécessaire pour connaître la probabilité d'un résultat statistique. Dans le cas de la méthode pour deux échantillons indépendants, il y a sept valeurs libres de varier dans chaque échantillon une fois que la moyenne est fixée, puisqu'il y a huit sujets dans chaque échantillon. Le nombre de degrés de liberté pour deux échantillons indépendants est fourni par l'équation suivante :

$$dl = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2 = 8 + 8 - 2 = 14 \tag{2.7}$$

Lorsque la méthode pour échantillons pairés est employée, c'est le nombre n de paires qui est pris en considération. La moyenne des différences étant fixée, le nombre de différences libres de varier est fourni par le nombre de paires moins 1, tel que calculé dans l'équation suivante :

$$dl = n - 1 = 8 - 1 = 7 \tag{2.8}$$

Une fois calculées les valeurs de t et de *dl*, il est possible de connaître la probabilité que les moyennes des deux groupes proviennent de la même population en consultant les tables de probabilités des manuels de statistique. Ces deux valeurs, celles de t et de *dl*, sont nécessaires pour juger de la probabilité de telles différences. Plus t est élevé, plus l'écart entre les échantillons est grand. Plus *dl* est élevé, plus la valeur de t peut être considérée comme représentative, puisqu'elle repose sur de grands échantillons. C'est pourquoi plus t et *dl* sont élevés, plus la probabilité que les moyennes de deux groupes proviennent de la même population est faible.

2.1 Types d'erreur en inférence statistique

Prenons maintenant un exemple tiré de la pratique psychologique. Supposons que l'on vous demande de prédire quels élèves âgés de 14 à 18 ans d'une école risquent de commettre une tentative de suicide au cours des trois prochaines années. Vous consultez les statistiques nationales et vous apprenez que, chaque année, 4 jeunes de

cette population sur 10 000 attentent à leur vie. Sur cette base, vous pourriez prédire qu'un jeune se suicidera ou ne se suicidera pas. Si, par exemple, vous rencontrez 2 500 élèves et que vous prédisez à chacun qu'il ne se suicidera pas, votre prédiction sera beaucoup plus souvent exacte qu'inexacte. En fait, vous avez 4 chances sur 10 000 de vous tromper, ce qui correspond à 1 chance sur 2 500.

Il y a dans cet exemple deux types d'erreur possible : (1) vous pouvez déclarer qu'un élève qui n'est pas suicidaire risquera d'attenter à sa vie ; (2) vous pouvez déclarer qu'un élève suicidaire n'attentera pas à sa vie. Dans ce cas-ci, comme dans bien des cas que nous rencontrons en statistiques également, les deux types d'erreur n'ont pas la même importance. L'erreur consistant à ne pas prédire qu'un élève suicidaire attentera à sa vie a de plus graves conséquences que l'erreur consistant à prédire qu'un élève qui n'est pas suicidaire attentera à sa vie.

Il y aurait peu d'intérêt à développer un outil de dépistage de prévention du suicide chez la population des 14-18 ans, la probabilité d'un tel événement étant déjà tellement faible qu'il serait peu probable qu'un tel outil fasse mieux qu'une prédiction "nulle", c'est-à-dire prédire que tous ces sujets ne se suicideront pas. Par contre, si l'on pouvait démontrer qu'un jeune sur deux âgé de 14 à 18 ans, ayant décroché de l'école, s'étant retrouvé sans emploi et ayant eu des antécédents d'alcoolisme ou de dépendance narcotique, risque d'effectuer une tentative de suicide, alors la mise au point d'un tel outil de dépistage pourrait être profitable, car, sur la seule base du hasard, nous aurions une chance sur deux (50%) de faire une prédiction exacte.

Ce sont de tels éléments de probabilité que les compagnies d'assurance utilisent pour le calcul des primes d'assurance automobile. Par exemple, un sujet célibataire, fumeur, de sexe masculin, de moins de 18 ans, conduisant une voiture sport, travaillant à plus de 15 km de son domicile et ayant des antécédents de mauvaise conduite constitue un risque plus grand que la moyenne générale des conducteurs. Ce risque est pris en compte dans le calcul des primes individuelles. Ceci ne veut pas dire que ce sujet fera inévitablement un accident, mais qu'il fait partie d'un groupe où le risque est plus grand que dans la population générale.

Tout comme l'actuaire, le chercheur scientifique qui se sert des statistiques pour formuler une conclusion, doit soupeser les probabilités associées à différents risques d'erreur. Tout comme nous l'avons vu dans l'exemple de la prédiction du risque de suicide, il existe deux types d'erreur en inférence statistique :

- *l'erreur de type I* consiste à affirmer, sur la base de probabilités extrêmement faibles, qu'un événement ne s'est pas produit au hasard, alors que de fait, un événement extrêmement rare, mais possible, vient de se produire.
- *l'erreur de type II* consiste à affirmer, sur la base de probabilités obtenues, qu'un événement a toutes les chances de s'être produit au hasard, alors que de fait, cet événement est le résultat d'un effet expérimental non négligeable.

Voyons maintenant comment ces deux types d'erreur s'appliquent à un cas concret tel que celui du test t de comparaison de deux moyennes. Après avoir comparé la moyenne de deux groupes de 25 élèves à un examen de mathématiques, un praticien calcule une valeur de t égale à 3,1 ($dl = 48$), ce qui d'après les tables de probabilités de la loi t de Student, se produit moins d'une fois sur 100. Deux interprétations s'offrent alors au praticien :

1. affirmer que les deux groupes ne sont pas différents quant à leur rendement en mathématiques et que l'écart observé résulte d'un effet du hasard qui se produit moins d'une fois sur 100 ;
2. affirmer que les deux groupes sont différents quant à leur rendement en mathématiques et que l'écart observé résulte d'un effet autre que le hasard.

En ce qui concerne la première hypothèse, appelée *hypothèse nulle* (H_0), il sera très difficile de contredire les personnes qui feront valoir qu'il est très peu probable que les groupes soient semblables. Considérant qu'un écart tel que celui observé ne se produit au hasard qu'une fois sur 100, il faudrait avoir été bien malchanceux pour tomber précisément sur cette possibilité. Il est plus cohérent de considérer qu'il existe une réelle différence entre les deux groupes et d'admettre l'autre hypothèse, que nous appelons *hypothèse alternative* (H_1).

Il se peut cependant que l'hypothèse nulle soit, malgré tout, correcte. C'est le cas chaque fois qu'un tel écart se produit effectivement au hasard, soit une fois sur 100 : c'est le deuxième type d'erreur. À première vue, cette alternative est peu défendable. Le risque de l'erreur associé à l'acceptation de l'hypothèse nulle (erreur de type II) semble en effet beaucoup plus grand que l'erreur associée à l'acceptation de l'hypothèse alternative (erreur de type I). Mais si le praticien vous informait que sur les quelques 80 tests de mathématiques administrés aux deux groupes depuis le début de l'année, c'est la première fois qu'un tel écart se manifeste, l'acceptation de l'hypothèse nulle pourrait être défendable.

2.2 PRISE DE DÉCISION STATISTIQUE ET NIVEAU DE SIGNIFICATION

La prise de décision statistique fait intervenir plusieurs facteurs. Il existe toujours un certain degré d'incertitude qui dépend de ce que nous considérons comme un risque acceptable ou non. En effet, quel pourcentage des différences s'étant produites au hasard entre deux groupes sommes-nous prêts à considérer comme extrêmes au point de nous faire préférer l'hypothèse alternative comme explicative des résultats ?

Dans la pratique, certains chercheurs opteront pour des pourcentages, appelés *niveaux de signification*, de l'ordre de 5% et moins. Ce pourra être 5%, 1% ou même 0,1% (respectivement 0,05, 0,01 et 0,001). Le choix d'un niveau de signification dépend directement du risque d'erreur de type I que nous sommes prêts à tolérer : c'est-à-dire, la probabilité de rejeter l'hypothèse de non différence (hypothèse nulle), alors qu'elle est vraie. Ce degré de tolérance nous est en partie dicté par des considérations scientifiques et pratiques.

Quels facteurs entrent en jeu dans le choix d'un niveau de signification plutôt qu'un autre ? Un chercheur qui en est à la phase exploratoire d'un programme de recherche ne voudra pas commettre l'erreur qui consiste à déclarer non significative une différence même petite. Il cherchera à réduire l'erreur de type II et pour cela, il choisira un niveau de signification plus grand, tel que 0,05. Parce qu'il ne veut pas fermer la porte à des différences qui, même petites, présentent un potentiel de recherche, il acceptera donc comme significatifs un plus grand nombre d'événements statistiques parmi les moins fréquents que s'il avait choisi un niveau de signification tel que 0,01 ou 0,001. Par contre, avant de déclarer qu'il existe des différences entre individus de

races différentes, il voudra s'assurer qu'il n'est pas tombé par hasard sur une différence inhabituellement grande. Dans de telles circonstances, étant donné l'importance et les répercussions qu'auront ses conclusions, le chercheur choisira de réduire l'erreur de type I en choisissant des niveaux de signification tels que 0,01 ou mieux encore 0,001. Il y en effet un risque important à déclarer que deux races sont différentes quant à une certaine caractéristique, alors qu'un écart tel que celui observé pourrait se produire au hasard 5 fois sur 100 entre deux groupes pour lesquels il n'existe pas de différence. Plus les conséquences de rejeter l'hypothèse nulle sont graves, plus le chercheur voudra se prémunir d'une erreur en adoptant un niveau de signification sévère (0,01 ou 0,001). Par contre, si c'est l'acceptation de l'hypothèse nulle qui constitue le plus grand risque, tel que de déclarer qu'une variable est sans effet alors qu'elle l'est réellement, alors le chercheur optera pour des niveaux tels que 0,05 et même 0,10.

2.3 PUISSANCE STATISTIQUE APPLIQUÉE À LA COMPARAISON DE DEUX MOYENNES

Plusieurs facteurs affectent la validité de notre prise de décision. Dans le cas de la comparaison de deux moyennes, l'un de ces facteurs a trait à la taille des échantillons. Plus les échantillons sont grands, plus nous nous attendons à ce que les moyennes soient similaires et plus nous serons portés à déclarer significatifs de faibles écarts. Un autre facteur a trait au risque que nous sommes prêts à prendre. Puisque nos décisions se fondent sur la probabilité que se produisent les différences observées, nous serons plus facilement enclins à déclarer significatifs des écarts lorsque nous acceptons une erreur de type I plus élevée. Enfin, le dernier facteur a trait à la méthode de calcul de la différence entre les moyennes.

Tableau 2 – Puissance et risques d'erreur associés à la décision statistique

		Situation dans la population	
		Hypothèse nulle vraie	Hypothèse nulle fausse
Décision statistique	Rejeter l'hypothèse nulle	*Type I d'erreur* $p = \alpha$	*Décision correcte* $p = 1 - \beta$ = puissance
	Ne pas rejeter l'hypothèse nulle	*Décision correcte* $p = 1 - \alpha$	*Type II d'erreur* $p = \beta$

Le tableau 2 résume les notions d'inférence statistique décrites dans la section précédente. On y retrouve les types I et II d'erreur ainsi qu'un nouveau concept, celui de la *puissance statistique*. En effet, même si certains risques sont associés à la prise de décision statistique et qu'aucune certitude n'existe à ce sujet, la probabilité d'en arriver à la bonne décision varie selon les situations. C'est ainsi que lorsque la probabilité de prendre la bonne décision est très faible, il peut être inutile d'entreprendre la recherche. Cette probabilité de prendre la bonne décision est ce que nous appelons la *puissance statistique*.

La puissance statistique est intimement liée au risque d'erreur. Le tableau 2 indique que le type I d'erreur se produit lorsque l'on rejette l'hypothèse nulle à partir des données de notre échantillon alors que l'hypothèse nulle est vraie dans la population. La probabilité de commettre l'erreur de type I est égale au niveau de signification choisi au départ, soit α. Quant à l'erreur de type II, elle consiste à prendre la décision de ne pas rejeter l'hypothèse nulle, alors qu'elle est fausse dans la population. La probabilité de l'erreur de type II nous est donnée par β. La complémentaire de l'erreur de type II, $1-\beta$, nous donne la probabilité de rejeter l'hypothèse nulle lorsqu'elle est fausse, ce qui constitue la puissance statistique d'un test. C'est pourquoi nous retrouvons toujours les valeurs de β dans les tables statistiques associées au calcul de la puissance d'un test.

Malheureusement, il est impossible, sans changer les conditions expérimentales, de minimiser à la fois les risques d'erreur de type I et de type II. Si l'on diminue la probabilité d'une erreur de type I, l'on accroît la probabilité de commettre une erreur de type II et réciproquement. Comment faire pour réduire simultanément les deux types d'erreur et, par conséquent, accroître la puissance de notre décision statistique ? Nous savons que plus l'échantillon est grand, meilleurs sont nos estimations des paramètres de la population. Par conséquent, nous pouvons parvenir à un meilleur test d'hypothèse en augmentant la taille des échantillons.

Une autre façon d'accroître la puissance d'un test consiste à utiliser la technique statistique qui représente le meilleur modèle de la situation que nous voulons tester. Certains tests statistiques sont mieux adaptés pour mettre à l'épreuve certaines hypothèses. C'est ce que nous avons vu avec l'exemple présenté dans le tableau 1. Dans cet exemple, nous avons testé l'hypothèse nulle qu'il n'existe aucune différence entre deux moyennes en utilisant deux tests statistiques différents : le test *t* pour deux échantillons indépendants et le test t pour échantillons pairés. Alors que l'écart entre les moyennes demeure le même dans chacun des cas, la valeur de *t* et la probabilité qui lui est associée varie. Dans le cas du test *t* pour deux échantillons indépendants, la probabilité associée à la valeur de *t* (0,26) est bien supérieure au niveau de signification que nous exigeons habituellement pour rejeter l'hypothèse nulle. Cette probabilité indique qu'une valeur de *t* comme celle que nous avons obtenu a 26 chances sur 100 de se produire au hasard, ce qui ne constitue pas un événement suffisamment rare pour que nous acceptions l'hypothèse alternative. Par contre, dans le cas du test pour deux échantillons pairés, la probabilité associée à la valeur de *t* (0,03) est telle que nous sommes conduits à accepter l'hypothèse alternative, puisque la probabilité qu'une telle valeur se produise n'est que de 3 sur 100. Comme nous étions prêts à déclarer significatifs des événements statistiques qui se produisent 5 fois sur 100 et moins, nous rejetons l'hypothèse nulle en faveur de l'hypothèse alternative.

Comment expliquer de tels écarts entre les résultats de ces deux tests statistiques, alors que les moyennes des deux groupes sont les mêmes ? La réponse réside dans la façon dont la procédure statistique traite les résultats. Dans le cas du test *t* pour échantillons indépendants, il n'est pas possible de comparer chaque sujet à un sujet bien précis de l'autre groupe puisqu'il n'existe aucune raison valable d'associer un sujet d'un groupe avec un sujet de l'autre groupe. La comparaison est donc globale et le test *t* porte sur l'écart des moyennes des deux groupes. Dans le cas du test *t* pour échantillons pairés, il existe un tel rationnel. La comparaison est donc spécifique et le

test t porte sur la moyenne des écarts observés entre chaque paire. Plus le pairage est efficace, plus la variable externe associée au pairage est importante dans l'explication des différences entre les résultats des deux sujets, plus le test t pour échantillons pairés est puissant par rapport au test t pour deux échantillons indépendants.

L'observation des données pour deux échantillons pairés indique que, même si les données sont les mêmes que pour deux échantillons indépendants, elles ont été réorganisées par paires. Le pairage démontre également que l'individu le plus faible du groupe 1 est généralement le plus faible dans le groupe 2, et que le plus fort dans le groupe 1 est le plus fort dans le groupe 2. Les deux échantillons sont liés et le pairage a donc réussi (nous pourrions dire également que les échantillons sont *corrélés*). Bien que la moyenne des différences et la différence des moyennes soient identiques pour chaque méthode (écart = -1,5), la valeur de t passe de 0,95 (dl = 16) à 2,29 (dl = 7) dans le cas de deux échantillons pairés. La probabilité que ces deux échantillons proviennent de la même population passe de 0,26 à une valeur beaucoup plus faible, soit 0,03. Il y a donc un lien entre les deux groupes qui s'explique par l'effet du pairage. Cet effet du pairage fait que le test t pour échantillons pairés est un modèle plus adéquat pour traiter les données. Un chercheur qui aurait traité les données de ces deux échantillons pairés au moyen d'un test pour deux échantillons indépendants, n'aurait pas rejeter l'hypothèse nulle alors qu'elle est fausse. Il aurait ainsi commis une erreur de type II à cause d'un test statistique moins puissant.

Que se produirait-il si le pairage n'avait aucun effet ? Si nous avions pairé les sujets en fonction de leur taille, le test t pour échantillons pairés n'aurait pas été plus puissant. En l'absence d'une variable adéquate de pairage, c'est le modèle pour échantillons indépendants qui convient le mieux.

3. Comparaison de plus de deux moyennes

Lorsque nous devons comparer plus de deux moyennes, le problème de la comparaison se pose différemment. Il arrive fréquemment que nous voulions savoir si k échantillons sont tirés de la même population ou si au moins l'un d'entre eux peut être considéré comme provenant d'une population différente. La tentation est forte d'utiliser le test t que nous venons de décrire en multipliant les comparaisons. Dans le cas d'un test impliquant cinq groupes, le nombre possible de tests t serait égal au nombre de combinaisons de deux dans cinq soit :

$$C_k^N = \frac{N}{k!\,(N-k)\,!} = C_2^5 = \frac{5!}{2!\,(5-2)\,!} = \frac{5\times4\times3!}{2!3!} = 10 \qquad (2.9)$$

En plus d'être peu pratique, une telle façon de procéder accroît considérablement les chances de déclarer significatives des différences occasionnées par les fluctuations d'échantillonnage, puisque nous effectuons 10 comparaisons de moyennes, chacune avec un risque d'erreur de type I égale au niveau de signification *par comparaison*. Mises ensemble, ces erreurs de type I dépassent ce qui est normalement accepté en inférence statistique pour prendre la décision d'accepter ou de rejeter l'hypothèse nulle.

3.1 Comparaisons multiples et taux d'erreur

Les comparaisons multiples entraînent deux taux d'erreur :

1. le taux par expérience (*experimentwise error rate*) ;
2. le taux par famille de comparaisons (*familywise error rate*).

Le premier se produit lorsque nous effectuons plusieurs comparaisons à partir de données recueillies sur les mêmes échantillons. Chacune de ces comparaisons ne peut être considérée comme indépendante des autres puisque les mêmes échantillons sont employés à chaque fois. C'est le cas lorsque nous comparons les moyennes des garçons et des filles pour chacune des 50 questions comprises dans un questionnaire. Pour l'ensemble de ces comparaisons, le taux par expérience est beaucoup plus élevé que le taux choisi par comparaison. Si le risque d'erreur par comparaison a été fixé à 0,05, le taux pour l'ensemble de cette expérience sera c fois plus grand, tel que calculé dans l'équation suivante :

$$\alpha = c\alpha' = 50 \times 0{,}05 = 2{,}5 \qquad (2.10)$$

Un tel taux d'erreur indique que parmi les 50 comparaisons, la probabilité est très forte que deux ou trois tests statistiques donneront lieu à une erreur de type I. Par conséquent, le chercheur déclarera significatives des différences produites par les fluctuations d'échantillonnage.

Parfois, nous sommes intéressés non pas à réaliser toutes les comparaisons possibles, mais une famille de comparaisons indépendantes entre elles. C'est le cas, lorsqu'en comparant les moyennes de cinq groupes, nous choisissons celles qui ont un intérêt particulier pour notre étude. Si, le groupe 5 est le groupe contrôle et que les quatre autres groupes constituent autant de groupes expérimentaux, il se peut que quatre comparaisons nous intéressent vraiment : celles entre les quatre groupes expérimentaux et le groupe contrôle. Ces quatre comparaisons sont indépendantes et le taux d'erreur pour cette famille de comparaisons se calcule différemment du taux par expérience. Il est donné par l'équation suivante :

$$\alpha = 1 - (1 - \alpha')^{c} = 1 - (1 - (0{,}05))^{4} = 0{,}1855 \qquad (2.11)$$

Le taux calculé (0,1855) pour l'ensemble des quatre comparaisons est bien supérieur au risque d'erreur de type I pour chacune des comparaisons $(\alpha' = 0{,}05)$. Le caractère cumulatif du risque d'erreur doit donc être pris en considération lorsque nous multiplions les tests de comparaison.

3.2 Analyse de variance et calcul du rapport *F*

Pour éviter, au moyen de comparaisons multiples, d'accroître l'erreur de type I, nous avons besoin d'un test d'hypothèse qui nous permette d'effectuer, en une seule fois, une comparaison de plusieurs moyennes. L'analyse de variance permet un test simple de l'hypothèse selon laquelle k échantillons ont été tirés d'une même population ou de populations équivalentes. Comme son nom l'indique, cette technique statistique met à profit l'analyse des différentes formes d'estimation de la variance afin de pouvoir confirmer ou infirmer cette hypothèse.

Le tableau 3 présente la simulation du tirage de cinq échantillons de 25 sujets tirés au hasard de la même population en ce qui concerne les quotients d'intelligence (moyenne = 100 ; écart-type = 15). Si les cinq échantillons ont été tirés de la même population, les différences entre les moyennes des cinq groupes devraient s'expliquer uniquement par les fluctuations d'échantillonnage. Mais, comment en être sûr ?

Tableau 3 – Simulation #1 : tirage de cinq échantillons de 25 sujets (μ=100 ; σ=15)

Données	Groupe 1	Groupe 2	Groupe 3	Groupe 4	Groupe 5
	56,77	112,35	85,78	128,90	100,40
	80,89	97,58	123,69	121,31	104,84
	76,94	119,50	102,99	104,99	86,73
	106,41	79,18	110,04	94,96	128,03
	106,28	115,87	86,52	109,05	114,24
	113,62	115,78	105,43	92,11	91,31
	85,42	116,86	113,69	104,87	128,16
	109,25	84,32	136,10	103,81	94,38
	92,48	79,03	71,42	90,79	86,02
	98,14	97,13	101,00	110,43	94,39
	114,03	106,69	107,57	123,02	107,95
	86,81	116,29	67,09	111,18	104,56
	113,29	105,71	99,85	102,13	106,83
	99,57	59,67	103,00	111,47	89,96
	113,46	115,75	62,25	100,35	117,33
	76,50	107,87	98,01	110,20	96,06
	103,38	110,49	130,00	87,93	57,56
	100,02	104,20	110,78	110,71	101,08
	80,13	96,20	101,20	85,16	97,66
	110,51	120,02	109,24	103,41	142,82
	84,53	105,08	90,48	126,09	97,96
	108,26	90,86	91,08	87,27	96,87
	126,55	77,26	100,18	98,93	92,57
	100,19	91,84	111,22	104,52	103,34
	76,44	123,17	92,87	124,75	134,40

En inspectant les statistiques descriptives des résultats des cinq groupes, il est difficile de se prononcer sur l'existence d'une différence quelconque entre les moyennes. Le tableau 4 indique que le groupe 1 est celui dont la moyenne est la plus basse (96,79) et le groupe 4, celui dont la moyenne est la plus élevée (105,93). À l'exception de ces deux valeurs extrêmes, les moyennes des autres groupes gravitent autour de la valeur de la moyenne de la population. Pour nous prononcer sur l'existence d'une différence entre une ou plusieurs des moyennes, il faudrait déterminer si les écarts observés entre les moyennes des cinq groupes sont le résultat de fluctuations normales d'échantillonnage. Bref, il nous faudrait connaître la probabilité de tirer au hasard cinq moyennes telles que celles que nous avons tirées.

Tableau 4 – Statiques descriptives du tableau 3

Groupe	n_i	Somme	Moyenne	Variance	Écart-type
Groupe 1	25	2419,8719	96,794877	270,77549	16,455257
Groupe 2	25	2548,7107	101,94843	267,00095	16,340164
Groupe 3	25	2511,4953	100,45981	308,99271	17,578189
Groupe 4	25	2648,3346	105,93338	154,15413	12,415882
Groupe 5	25	2575,4547	103,01819	314,13254	17,723784
Moyenne			101,63094	263,01116	16,217619
Variance des moyennes			11,323831		

ANOVA

Source	Sc	dl	Nc	F	Prob.deF	Valeur crit. F
Inter Groupe	1132,3831	4	283,095779	1,0763641	0,3713259	2,44723708
Intra Groupe	31561,34	120	263,01116			

Nous disposons déjà d'un moyen simple de déterminer le degré de variation possible entres les moyennes tirées d'une même population. Dans la section 1.2, portant sur l'estimation de la moyenne d'une population, nous avons vu que la variance des moyennes était n fois plus petite que celle des résultats, n représentant la taille de l'échantillon. En effet, plus les moyennes de chaque groupe sont calculées à partir d'échantillons de grande taille, plus petite devrait être leur variation. Est-ce bien le cas dans l'exemple du tableau 4 ?

Pour calculer la variance de moyennes, nous procédons de la même manière que pour la variance des résultats. Voici un exemple de calcul à partir des données du tableau 4 :

$$s_{moy}^{2} = \sum \frac{(\bar{X} - \bar{M})}{k} = \frac{(96{,}79 - 101{,}63) + \ldots + (103{,}02 - 101{,}63)}{5} = 11{,}32 \qquad (2.12)$$

où $\bar{M}$ représente la moyenne des moyennes de chaque groupe et k = nombre de moyennes (ou de groupes).

La variance des moyennes est bien inférieure à n'importe quelle variance des résultats observée pour chacun des cinq groupes. En effet, la variance des résultats s'étend de 154,15 pour le groupe 4 jusqu'à 308,99 pour le groupe 3. Selon ce que nous savons des lois d'estimation de la moyenne, la variance des moyennes devrait être 25 fois plus petite que la variance des résultats. Or, dans le cas du groupe 4, elle est 15 fois plus petite, alors que dans le cas du groupe 3, elle est environ 30 fois plus petite. Quelle devrait être notre décision ?

Notre serions sans doute mieux renseignés si, au lieu de comparer la variance des moyennes à la variance des résultats de chaque groupe, nous utilisions les résultats de tous les groupes pour calculer la variance des résultats. C'est ce que nous avons fait en calculant *la moyenne des variances* pour les cinq groupes, ce qui nous a donné 263,01. Il est normal que les variances des résultats de chaque groupe, même lorsque ces groupes sont tirés de la même population, ne soient pas identiques. La moyenne des variances nous fournit donc une meilleure estimation de la variance des résultats dans la population que ne pourrait le faire un seul groupe à la fois.

Nous pouvons donc comparer deux estimations de la variance des résultats de la population. L'une calculée à partir de la variance des moyennes que nous savons être n fois plus petite que la variance des résultats. L'autre calculée à partir de la moyenne de la variance des résultats, que nous savons être la meilleure estimation possible de la variance des résultats dans la population. Or, si les cinq groupes en présence ont été tirés de la même population (ou de populations aux caractéristiques identiques), il ne devrait pas y avoir de différences remarquables entre ces deux estimations.

Dans l'exemple du tableau 4, on peut estimer la variance des résultats de la population à partir de la variance des moyennes en utilisant la formule (2.3). Dans ce cas-ci, nous chercherons à résoudre cette équation non pas pour $s_{\bar{X}}^{2}$, mais pour s_{X}^{2}. En substituant par leurs valeurs respectives nous obtenons :

$$s_{X}^{2} = ns_{\bar{X}}^{2} = 25 \times 11{,}324 = 283{,}10 \qquad (2.13)$$

La variance des moyennes étant 25 fois plus petite que celle des résultats, nous pouvons estimer que la variance des résultats devrait être 283,10. Nous appelons *variance inter-groupes* ou *variance inter*, la variance des résultats de la population estimée de cette manière. Nous appelons *variance intra-groupes* ou *variance intra*, la variance des résultats estimée en calculant la moyenne des variances de chacun des groupes. Nous savons que celle-ci vaut 263,01, tel qu'indiqué dans le tableau 4. Cette valeur est simplement la moyenne des variances des cinq groupes :

$$\frac{270{,}78 + 267{,}00 + 308{,}99 + 154{,}15 + 314{,}13}{5} = 263{,}01 \qquad (2.14)$$

La comparaison de ces deux valeurs confirme que la variance des moyennes n'est pas inhabituelle. En effet, si nous faisons le rapport — appelé *F* d'après le nom de l'initiateur de cette méthode, le statisticien Fisher — entre les deux valeurs estimées de la variance des résultats de la population, la *variance inter* et la *variance intra*, nous obtenons une valeur voisine de 1 :

$$F = \frac{\text{Variance inter}}{\text{Variance intra}} = \frac{283,10}{263,01} = 1,08 \tag{2.15}$$

Un rapport $F = 1$ indique que les deux estimations sont égales. Si le rapport *F* calculé entre les deux estimations de la variance des résultats dans la population n'est pas très différent de 1, alors nous avons de bonnes raisons de croire que les écarts entre les moyennes sont purement aléatoires et que tous les groupes en présence peuvent être considérés comme ayant été tirés de la même population. Pour en être vraiment convaincu, il faudrait connaître de façon précise la probabilité d'obtenir la valeur observée de *F* ou une valeur plus extrême, *lorsque l'hypothèse nulle est vraie*. Nous aborderons cette question lorsque nous parlerons de la loi des probabilités de *F*.

Voyons maintenant ce qui se passerait si certains des groupes tirés au hasard ne provenaient pas de la même population. C'est ce que nous avons tenté de simuler dans le tableau 5. Pour réaliser cette simulation, nous avons soustrait 3 de tous les résultats du groupe 1 et nous avons additionné 5 à tous les résultats du groupe 4. Ces valeurs correspondent à un effet expérimental qui pourrait se produire si, dans le cas des résultats de QI, nous avions tiré notre échantillon de populations différentes : par exemple, une population d'étudiants ayant terminé sa scolarité obligatoire (+5) et une population d'étudiants ne l'ayant pas terminée (-3).

Tableau 5 – Simulation #2. Effets expérimentaux : Groupe 1 = (-3) ; Groupe 4 = (+5)

Groupe	n_i	Somme	Moyenne	Variance
(Groupe 1) -3	25	2344,871934	93,79487737	270,7754878
Groupe 2	25	2548,710735	101,9484294	267,0009469
Groupe 3	25	2511,495304	100,4598121	308,9927133
(Groupe 4) +5	25	2773,334602	110,9333841	154,1541329
Groupe 5	25	2575,454718	103,0181887	314,1325352
Moyenne			102,0309383	263,0111632
Variance des moyennes			37,63403733	

Source	sc	dl	MC	F	Prob. de F	Valeur crit. F
Inter groupes	3763,40373	4	940,8509326	3,577228134	0,008577502	2,44723708
Intra groupe	31561,33959	120	263,0111632			
Total	35324,74332	124				

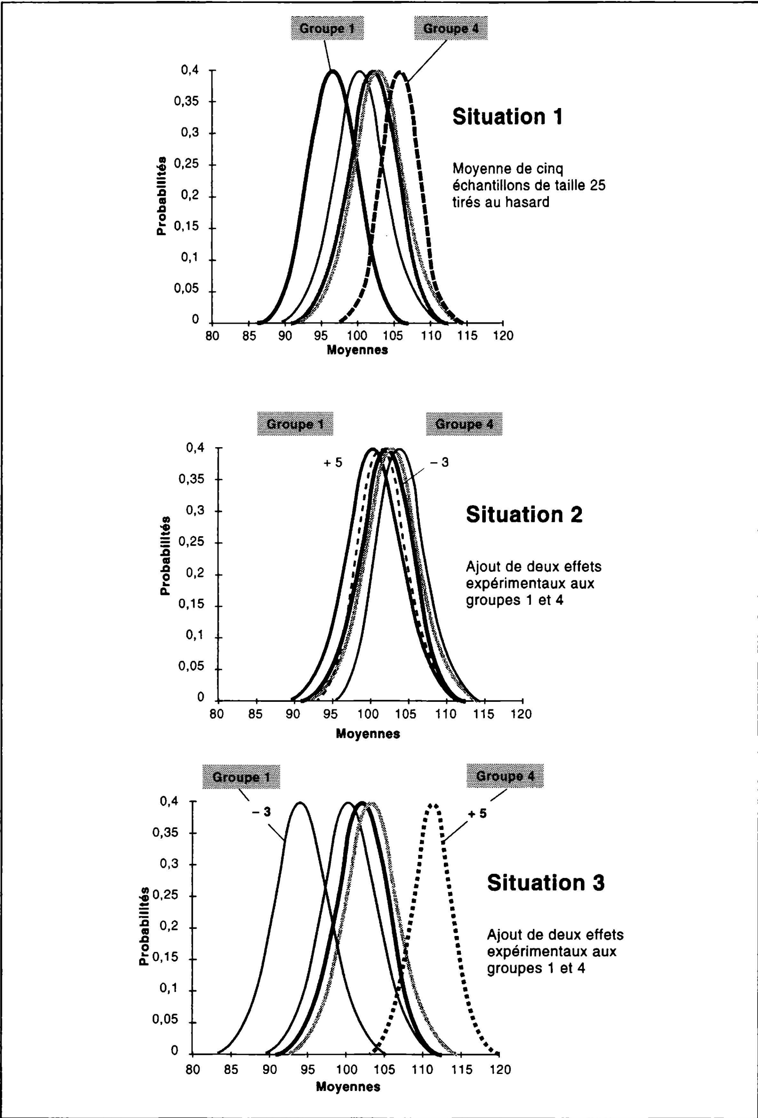


Figure 3 – Représentation graphique de trois situations d'ANOVA (distributions des moyennes de chaque groupe et erreurs d'estimation)

Comme l'illustre la figure 5, l'addition de ces effets expérimentaux a eu pour effet d'éloigner les groupes 1 et 4 des autres groupes situés plus près de la moyenne générale de la population. Mais, cet écart est-il suffisant pour être déclaré significatif ? Pour répondre à cette question, il faut calculer la probabilité que de telles différences se produisent au hasard

En ajoutant deux effets expérimentaux aux groupes 1 et 4, nous avons changé la variance entre les moyennes des cinq groupes. C'est ce que nous révèle la figure 3 (situation #3), et c'est aussi ce que nous retrouvons en consultant le tableau 5 pour la valeur calculée de la variance des moyennes. Celle-ci est maintenant de 37,63 (au lieu de 11,32), ce qui traduit bien les conséquences des effets expérimentaux. Par contre, en ajoutant +5 et -3 aux résultats des groupes 1 et 4, la variance des résultats pour chacun des groupes n'a pas changé. Il en est de même lorsque nous calculons la moyenne des variances des cinq groupes : celle-ci demeure inchangée par rapport à la situation initiale où nous n'avions ajouté aucun effet expérimental.

L'addition d'effets expérimentaux aux résultats de deux des cinq groupes n'a pas les mêmes conséquences sur l'estimation de la variance de la population, que celle-ci s'effectue à partir de la variance des moyennes (variance inter) ou à partir de la moyenne des variances des résultats de chaque groupe (variance intra). Lorsque nous avons ajouté un effet expérimental de +5 au groupe 4, la moyenne s'est trouvée accrue de la même valeur, mais la variance des résultats est demeurée inchangée. Le même phénomène s'est produit en ajoutant -3 au groupe 1.

La variance intra, calculée à partir de la moyenne des variances à l'intérieur de chaque groupe, n'est pas affectée par les effets expérimentaux. Elle constitue donc une estimation *non biaisée* de la variance de la population. Par contre, la variance des moyennes est affectée par ces effets expérimentaux et la variance inter (n fois la variance des moyennes) est donc une estimation *biaisée* de la variance de la population. En effet, si l'on cherche à estimer la variance des résultats de la population à partir de la variance entre les moyennes dans l'exemple du tableau 5, nous trouvons :

$$s_X^2 = Ns_{\overline{X}}^2 = 25 \times 37{,}63 = 940{,}85 \tag{2.16}$$

Cette valeur est plus de trois fois supérieure à celle de la variance des résultats de la population calculée à partir de la moyenne des variances de chaque groupe, tel que le démontre le calcul du rapport F :

$$F = \frac{\text{Variance inter}}{\text{Variance intra}} = \frac{940{,}85}{263{,}01} = 3{,}58 \tag{2.17}$$

L'ajout d'effets expérimentaux a provoqué une hausse importante du rapport F, faisant passer celui-ci d'une valeur voisine de 1, lorsque les seules variations sont dues aux effets d'échantillonnage, à une valeur de 3,58 lorsque nous avons ajouté des effets expérimentaux à deux des cinq groupes. F est donc un bon indicateur du degré de différence entre les moyennes. Il nous permet de déterminer si les fluctuations que nous observons entre les moyennes des groupes sont probables pour des échantillons tirés d'une population où il n'y a pas de différences (H_0 vraie) ou si elles sont peu probables. La probabilité associée à cet indicateur peut nous servir à prendre une décision quant à l'existence ou non d'une différence significative.

La décision prise à partir du rapport *F* peut être entachée d'erreur. Nous devons considérer le caractère particulier de la simulation précédente. La valeur de l'effet expérimental +5 a été ajoutée au groupe 4, dont la moyenne était déjà la plus élevée, et la valeur de l'effet expérimental -3 a été ajoutée au groupe 1, dont la moyenne était déjà la plus basse. Ceci a eu pour conséquence d'accroître les écarts entre les moyennes des groupes tirés au hasard.

Lorsque nous réalisons une recherche, les effets expérimentaux se distribuent au hasard. C'est ainsi que pour évaluer cinq méthodes d'apprentissage des mathématiques, nous choisissons 125 sujets que nous associons au hasard à chacune des cinq méthodes. Il n'y a pas de raison de suspecter que les individus de faible QI aient une probabilité plus grande d'être associés à la moins bonne des méthodes (-3) et que les sujets les plus intelligents soient associés à la meilleure (+5). L'effet le plus fort peut être attribué à n'importe quel groupe, tout comme l'effet le plus faible.

Nous pouvons envisager cependant une situation où l'effet +5 est ajouté au groupe le plus faible, alors que l'effet -3 est ajouté au groupe le plus fort (tableau 6). Les conséquences de cette simulation, illustrée dans la situation #2 de la figure 3, sont de rapprocher les moyennes les unes des autres et de réduire les écarts observés lors des fluctuations normales d'échantillonnage.

Tableau 6 – . Simulation #3. Effets expérimentaux : Groupe 1 = (+5) ; Groupe 4 = (-3)

Groupe	n_i	Somme	Moyenne	Variance
(Groupe 1) +5	25	2544,871934	101,7948774	270,7754878
Groupe 2	25	2548,710735	101,9484294	267,0009469
Groupe 3	25	2511,495304	100,4598121	308,9927133
(Groupe 4) -3	25	2573,334602	102,9333841	154,1541329
Groupe 5	25	2575,454718	103,0181887	314,1325352
Moyenne			102,0309383	263,0111632
Variance des moyennes			1,080010421	

Source	sc	dl	MC	F	Prob. de F	Valeur crit. F
Inter groupes	108,0010421	4	27,00026053	0,10265823	0,981370812	2,44723708
Intra groupe	31561,33959	120	263,0111632			
Total	31669,34063	124				

3.3 ÉCHANTILLONNAGE ET ANALYSE DE VARIANCE

La simulation décrite dans le tableau 6 présente une situation où la variance inter est plus petite que la variance intra. Le rapport F est inférieur à 1 ($F = 0{,}1$) ce qui indique que les écarts entre les moyennes ne sont que le dixième de ce que nous serions en droit d'attendre si elles avaient varié aléatoirement. Lorsque l'hypothèse nulle est vraie et que nos procédures d'échantillonnage sont adéquates, les variations d'échantillonnage n'entraînent que très rarement des valeurs de F très inférieures à 1. Lorsqu'elles se produisent, il faut s'interroger sur la valeur de notre dispositif d'échantillonnage ou notre méthode d'attribution des différents traitements expérimentaux.

D'autres procédures d'échantillonnage ont pour effet d'exagérer les écarts entre les moyennes. En éducation, de telles situations sont fréquentes. C'est le cas lorsqu'un *échantillonnage par grappes* (voir chapitre 7) est employé au lieu d'un échantillonnage aléatoire. Ceci se produirait si plutôt que de tirer au hasard les 125 sujets de l'ensemble de la population des élèves de cinquième année de la ville d'Ottawa, un chercheur avait choisi — pour des raisons pratiques — cinq classes de 25 sujets. Une fois qu'une classe est choisie, tous les élèves de cette classe deviennent sujets de l'étude. Dans ce cas-ci, il est possible que les élèves d'une même classe soient plus homogènes qu'un groupe de 25 élèves tirés de l'ensemble de la population. La variance intra risque donc d'être sous-estimée. De plus, les moyennes de chaque classe risquent de refléter le milieu socio-économique des écoles dans lesquelles elles sont situées. Les écarts entre les moyennes de classes provenant de milieux différents risquent donc d'être exagérés. La variance inter risque de surestimer la variance de la population. Les deux facteurs mis ensemble font qu'il est beaucoup plus facile, au moyen d'un échantillonnage par grappes, d'obtenir un rapport F élevé puisque la variance inter surestimera la variance de la population et la variance intra la sous-estimera. Lord (1959) a démontré qu'il fallait des échantillons de taille 12 à 30 fois plus grande pour réaliser avec un échantillonnage par grappes des estimations de la moyenne similaires à celles d'un échantillonnage aléatoire simple.

3.4 Postulats de l'analyse de variance

Ces dernières observations nous permettent d'énoncer un certain nombre de conditions garantissant une utilisation appropriée de l'analyse de variance. Ces postulats sont les suivants :

1. les échantillons sont tirés au hasard d'une population normale ;
2. les observations sont indépendantes entre elles ;
3. les variances de l'ensemble des échantillons sont homogènes.

Ces postulats vont de soi. Si les variances des échantillons sont trop différentes, la variance intra, calculée à partir de la moyenne des variances de chaque groupe, n'est plus une estimation fiable de la variance de la population. Si les observations ne sont pas indépendantes, comme dans le cas d'un échantillonnage par grappes, l'estimation des variances inter et intra devient biaisée. Enfin, les distributions des résultats doivent permettre de considérer que chaque groupe a été tiré d'une population normale. Il

serait difficile de comparer des moyennes provenant de distributions qui diffèrent entre elles par leur symétrie, leur kurtose, etc.

3.5 Loi de probabilité de *F*

Si tous les postulats de l'analyse de variance ont été respectés, alors les sources de variation de la valeur *F*, lorsque l'hypothèse nulle est vraie, se limitent à deux :

1. le nombre de groupes ;
2. le taille de l'échantillon de chaque groupe.

Plus le nombre de groupes est élevé, plus la variance inter s'appuie sur un grand échantillon de moyennes pour estimer la variance de la population. De la même façon, plus la taille des groupes est élevée, plus l'estimation de la variance intra sera précise. En conclusion, la probabilité de *F* dépend de deux valeurs de degrés de liberté : le nombre de moyennes des groupes libres de varier *(k-1)* et le nombre de résultats libres de varier à l'intérieur de chaque groupe *(n-1)*.

Pour connaître la valeur de probabilité de *F*, il faut consulter une *table de Fisher*. Cette table comporte deux entrées : la première pour les degrés de liberté de la variance inter, la seconde pour les degrés de liberté de la variance intra. Plus les degrés de liberté sont élevés, plus il est possible de déclarer une différence significative entre les moyennes à partir d'une petite valeur de *F* supérieure à 1. Dans de telles circonstances, en effet, les estimations des variances inter et intra sont les plus précises.

3.6 Lecture d'un tableau d'analyse de variance (ANOVA)

La présentation des résultats d'une analyse de variance suit certaines conventions qui en facilitent l'interprétation. Les tableaux 4 à 6 vous en fournissent des modèles. Dans tous ces tableaux, les résultats des calculs sont présentés en indiquant dans chaque colonne les renseignements suivants :

1. la source (variance inter ou intra) ;
2. *SC* : la somme des carrés des écarts à la moyenne ;
3. *dl* : les degrés de liberté ;
4. *MC* : la moyenne des carrés ou variance. Elle est calculée en divisant la somme des carrés par le nombre de degrés de liberté ;
5. le rapport *F* ;
6. la probabilité associée à *F* ;
7. la valeur critique de *F* pour le niveau de signification choisi au préalable.

Dans le tableau de l'ANOVA, les sommes des carrés *SC* ne nous intéressent pas vraiment. Elles servent principalement au calcul des moyennes de carrés *MC*, ces estimations de la variance essentielles au calcul du rapport *F*. Pour interpréter ce rapport *F*, nous devons connaître sa probabilité pour les valeurs de degrés de liberté en présence. Si cette probabilité est tellement faible qu'il y a peu de chances qu'un tel rapport *F* se produise lorsque les moyennes ont été tirées au hasard de la même population, alors nous préférons accepter l'hypothèse alternative selon laquelle au moins une des

moyennes n'est pas tirée de la même population. À partir d'ici, nous appliquons les mêmes principes de décision statistique que ceux que nous avons vus pour la comparaison de deux moyennes (loi *t de Student*).

Une autre façon d'évaluer F consiste, non pas à en connaître la probabilité exacte, mais à en comparer la valeur à une valeur seuil, appelée *valeur critique*, correspondant aux degrés de liberté et au niveau de signification (type I d'erreur) choisi au préalable. Dans le cas du tableau 5, la valeur critique de F pour un niveau de signification de 0,05 vaut 2,45. Toute valeur de F supérieure à 2,45 aura moins de 5% des chances de s'être produite au hasard du fait de simples fluctuations d'échantillonnage. Dans ce cas-ci, la valeur calculée de F (3,58) étant supérieure à la valeur critique, nous choisirons de rejeter l'hypothèse nulle et d'accepter l'hypothèse qu'au moins une des moyennes est différente ou ne provient pas de la même population.

3.7 Puissance de l'ANOVA

L'analyse de variance est le test le plus puissant de comparaison de moyennes, lorsque les postulats sont respectés et que le modèle statistique employé convient bien au plan d'observation. Tout comme dans le cas du test t, il existe une probabilité plus ou moins grande de prendre la bonne décision, soit de rejeter l'hypothèse nulle lorsqu'elle est fausse, selon la précision avec laquelle nous estimons les moyennes et selon l'importance des effets expérimentaux.

Dans le cas des simulations précédentes, nous avons vu qu'une conjonction de circonstances particulières avaient contribué, dans un cas (tableau 6 ; situation #2, figure 3) à accepter l'hypothèse nulle, alors que dans un autre cas (tableau 5 ; situation #3, figure 3), nous avions choisi de la rejeter, et ce pour des effets expérimentaux identiques. Dans un cas, les fluctuations d'échantillonnage se sont ajoutées au effets expérimentaux pour accroître les différences entre les moyennes, alors que dans l'autre cas, elles ont contribué à les atténuer. Ces deux simulations décrivent une situation où la puissance statistique pourrait être qualifiée de relativement faible, parce que la variation causée par les effets expérimentaux n'est pas beaucoup plus grande que les effets d'échantillonnage, du moins avec des échantillons de cette taille. Sans se livrer à des calculs importants, on peut dire que des effets de +10 et -15 pourraient difficilement passer inaperçus avec des échantillons de 25 sujets tirés de la population que nous avons définie au départ. Par contre, pour déceler des effets de +3 ou -2, il faudrait réduire considérablement la variance d'échantillonnage et le seul moyen de le faire serait d'accroître considérablement la taille des échantillons.

Pour avoir une idée exacte, non seulement de la probabilité d'une différence, mais aussi de son importance et de sa grandeur, de plus en plus de statisticiens calculent, en plus du rapport F, une valeur indiquant la grandeur de l'effet expérimental. Il existe plusieurs façons de calculer une telle valeur, mais nous nous limiterons à la plus simple, η^2 *(êta-carré),* calculée au moyen de l'équation suivante :

$$\eta^2 = \frac{SC_{total} - SC_{intra}}{SC_{total}} = \frac{SC_{intra}}{SC_{total}} \qquad (2.18)$$

Si l'on calcule la valeur de η^2 pour les trois simulations et que les nous comparons aux valeurs et probabilités de *F*, nous obtenons les résultats présentés au tableau 7 :

Tableau 7 – Valeurs et probabilités de F, ainsi que de η^2 pour les simulations 1 à 3 (tableaux 4 à 6)

	F	Probabilités de F	η^2
Simulation 1 (H_0 vraie)	1,08	0,371	0,035
Simulation 2	0,10	0,981	0,003
Simulation 3	3,58	0,009	0,107

Ce tableau nous indique que même lorsque *F* est significatif, l'importance de l'effet expérimental ne dépasse guère 10% de la somme totale des carrés. Il reviendra au chercheur de déterminer si un tel effet expérimental, même significatif, a une importance suffisante pour justifier de nouvelles recherches.

3.8 Autres considérations sur l'ANOVA

L'analyse de variance nous aura permis d'illustrer une autre facette de l'inférence statistique. En fait, l'ANOVA constitue une famille de tests statistiques qu'il serait impossible de décrire en un seul chapitre. Tout comme la loi *t* de Student permet de comparer deux moyennes tirées d'échantillons indépendants ou liés, la loi *F* de Fisher permet de mettre à l'épreuve des modèles expérimentaux beaucoup plus complexes que le plan simple que nous venons de décrire. Ici encore, plus le modèle expérimental est approprié, plus puissante est notre décision statistique.

De nombreuses considérations entourent l'utilisation appropriée de l'ANOVA. Les exemples présentés sont des simulations qui représentent des cas idéaux. La réalité est plus diversifiée. Les échantillons peuvent être de tailles inégales suite au désistement d'un ou plusieurs sujets. Les distributions des résultats peuvent s'écarter sensiblement d'une distribution normale. Chacun de ces cas particuliers requiert une solution que l'on pourra étudier dans les nombreux ouvrages traitant d'analyse de variance et d'inférence statistique.

Suite à une analyse de variance, le chercheur peut être intéressé à déterminer entre quelles moyennes les différences sont significatives. Des tests de comparaisons multiples des moyennes sont alors nécessaires pour tenir compte du taux d'erreur par famille. L'analyse de variance nous permet de déclarer s'il existe une ou plusieurs moyennes qui diffère des autres. Elle ne nous précise pas cependant entre quelles moyennes ces différences se produisent. C'est pourquoi des tests *post hoc* existent afin de préciser entre quelles moyennes les différences les plus significatives se sont produites. Lorsque le chercheur, de par la formulation de ses hypothèses de recherche, ne s'intéresse qu'à un nombre restreint de comparaisons bien déterminées, le recours à des tests plus puissants de comparaisons *a priori* est alors possible.

Il y aurait encore beaucoup à dire sur l'analyse de la variance. En mesure, elle joue un rôle particulier comme moyen de calculer l'importance de différentes sources de variation dans l'étude de la généralisabilité, une méthode de calcul de la fidélité pour des plans complexes d'observation. Grâce à cette introduction, nous pourrons aborder de manière plus compréhensible ce sujet dans la section 6 du chapitre 4.

4. Relations entre variables : corrélation et régression linéaire

4.1 Description de la relation entre deux variables

Les constructeurs et les utilisateurs de tests sont intéressés par les relations qui existent entre les scores obtenus par les mêmes sujets sur différentes variables. Ces relations sont particulièrement importantes lorsque l'on étudie la validité d'un test ou d'un questionnaire et lorsque l'on désire réaliser des prédictions à partir des résultats d'une ou de plusieurs épreuves. Par exemple, on peut évaluer la relation entre les scores d'un test d'admission à l'université et les résultats académiques en fin de première année. On peut également apprécier la relation entre un questionnaire de dépression et les évaluations faites par des cliniciens. Ou encore, on peut mesurer la relation entre l'âge des enfants et leurs scores à un test de vocabulaire. Dans tous ces cas, on se demande dans quelle mesure les différences observées sur une des variables se reflètent sur l'autre. Deux indicateurs sont fréquemment utilisés dans ce but : le *coefficient de corrélation* et la *droite de régression.* Dans cette section, nous introduisons ces deux concepts. Nous insistons particulièrement sur les principes essentiels qui doivent guider l'interprétation des corrélations et des fonctions de régression linéaire.

La relation entre deux variables peut être représentée de manière graphique au moyen d'un *diagramme cartésien.* Les résultats sur la première variable sont notés sur l'axe horizontal, appelé *abscisse* (ou axe des X), et ceux sur la seconde variable sont notés sur l'axe vertical, appelé *ordonnée* (ou axe des Y). Chaque sujet possède ainsi deux coordonnées qui sont ses scores pour les deux variables en question. À partir de ces coordonnées, il est possible de situer un sujet sous forme d'un point dans l'espace bi-dimensionnel constitué par les deux axes (plan cartésien). Dans la figure 4, nous avons indiqué les points représentant la position de trois sujets sur base de leurs scores à un test de français et à un test de mathématique.

Nous pouvons réaliser la même représentation graphique pour tous les sujets d'un échantillon. Nous obtiendrons ainsi un nuage de points dont la forme nous donne une première indication de la relation existant entre les deux variables étudiées. La figure 5 présente quatre nuages de points qui constituent autant de types de relation entre les variables. Le graphique (a) est l'exemple d'une relation positive entre variables. À une augmentation sur la variable X correspond une augmentation sur la variable Y. C'est le type de relation que l'on peut, par exemple, observer entre le QI et les résultats scolaires. Dans le cas présent, la relation n'est pas parfaite, ce qui n'est le cas que lorsque l'augmentation de Y est exactement proportionnelle à chaque augmentation de X. Toutefois, malgré la variabilité de la relation, nous pouvons constater que le

nuage de points tend à prendre la forme d'une droite. Pour cette raison, la relation entre deux variables est qualifiée de *linéaire*. Nous reviendrons plus loin sur cette notion lorsque nous expliquerons le concept de *régression*.

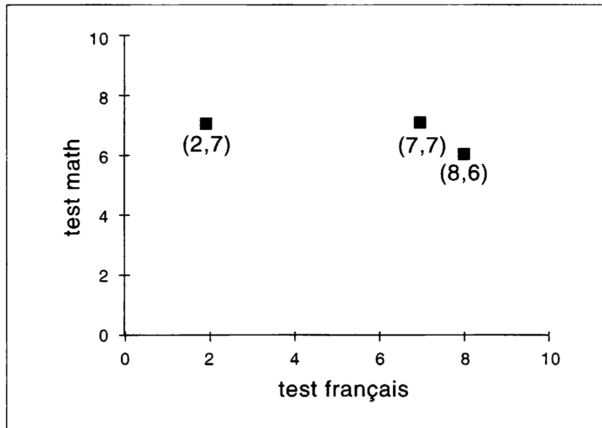


Figure 4 – Représentation graphique de la position de trois sujets en fonction de leurs scores à deux tests

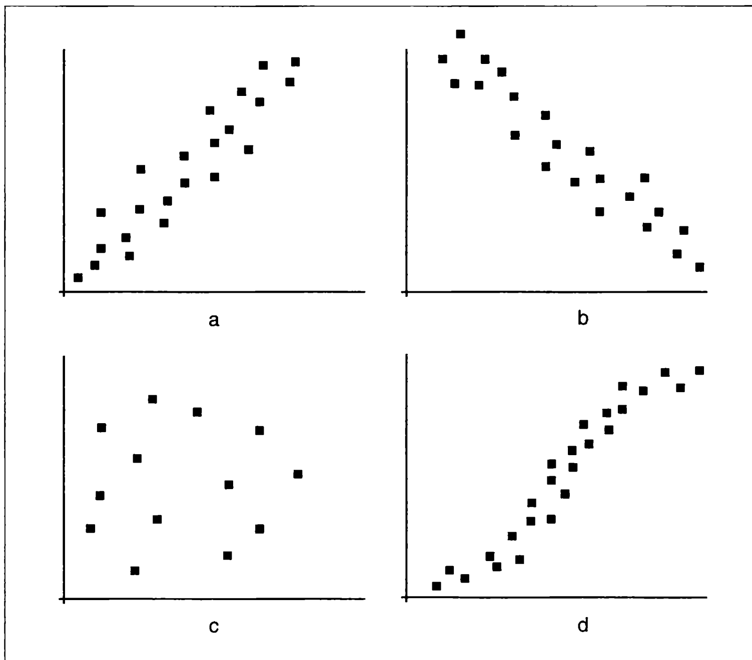


Figure 5 – Diagrammes cartésiens illustrant différents types de relations entre variables

Le graphique (b) illustre une relation négative entre les variables. Dans ce cas, à une augmentation de *X* correspond une diminution de *Y*. Nous pouvons observer une

telle relation lorsque, par exemple, nous comparons le vieillissement et les performances à un test de mémoire. Le graphique (c) est l'exemple d'une absence de relation entre les variables. Il n'y a aucune tendance systématique de *Y* à varier en même temps que *X*. Par conséquent, la valeur de *X* ne peut rien nous apprendre à propos de la valeur de *Y*. Enfin, le graphique (d) nous présente une relation non linéaire entre les variables. Il y a bien une relation entre *X* et *Y* mais celle-ci ne prend pas la forme d'une ligne droite. Dans l'exemple présent, le nuage de points prend la forme en S de l'ogive normale. Nous verrons dans le chapitre 8 différentes illustrations de ce type de relation dans le cadre des Modèles de la Réponse aux Items (MRI).

4.2 Le coefficient de corrélation

En plus d'une représentation graphique, il est possible quantifier la relation existant entre deux variables. Lorsque cette relation est fondamentalement linéaire et que les deux variables sont mesurées sur une échelle d'intervalle, on calcule habituellement le coefficient de corrélation de Bravais-Pearson. Celui-ci est égal à la covariance de *X* et de *Y* divisée par le produit des écart-types de *X* et de *Y* :

$$r = \frac{\mathrm{cov}_{XY}}{S_X S_Y} \tag{2.19}$$

Rappelons que la covariance de *X* et *Y* peut être calculée grâce à la formule suivante :

$$\mathrm{cov}_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n - 1} \tag{2.20}$$

Après développement, la formule permettant de calculer le coefficient de corrélation peut dès lors s'exprimer de la manière suivante :

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}} \tag{2.21}$$

Lorsque les deux distributions sont exprimées en scores *z*, et qu'elles ont donc une même moyenne égale à 0 et un même écart-type égale à 1, une formule beaucoup plus simple peut être utilisée :

$$r = \frac{\sum z_X z_Y}{n} \tag{2.22}$$

Le coefficient de corrélation peut varier de -1,00 à +1,00. Lorsqu'il est égal à +1,00, nous avons affaire à une corrélation positive parfaite entre les variables X et Y. Lorsqu'il est égal à -1,00, nous avons affaire à une corrélation négative parfaite entre ces deux variables. Lorsqu'il est égal à 0, les deux variables sont non corrélées. Nous pouvons illustrer l'usage de cette formule avec les données présentées dans le tableau 8. Il s'agit des résultats de deux tests passés par un échantillon de 92 enfants âgés de 8 ans et demi. Le premier test est une épreuve de calcul mental et le second évalue la mémoire de séries de chiffres. Pour des raisons de place, nous ne donnons ici qu'une

partie des données. Par contre, nous présentons tous les résultats des calculs intermédiaires qui permettent ensuite de calculer le coefficient de corrélation.

Tableau 8 – Résultats d'un test de calcul mental et d'un test de mémoire (N=92)

Sujets	Test de calcul	Test de mémoire
1	8	7
2	9	6
3	9	9
4	8	8
5	6	11
6	5	9
7	16	12
8	10	8
9	13	17
10	6	9
...	...	...

$$\sum X = 934 \quad \sum X^2 = 10430 \quad \sum Y = 941 \quad \sum Y^2 = 10503 \quad \sum XY = 10046 \quad (2.23)$$

$$r = \frac{92 \times (10046) - (934 \times 941)}{\sqrt{(92 \times (10430) - 872356)\,(92 \times (10503) - 885481)}} = 0{,}54 \quad (2.24)$$

Il ne suffit pas de calculer correctement le coefficient de corrélation encore faut-il l'interpréter adéquatement. Que signifie en effet une corrélation de 0,54 entre deux tests ? Pour réaliser cette interprétation, un certain nombre de règles doivent être respectées.

Il faut tout d'abord tenir compte de la signification statistique du coefficient obtenu. Celui-ci est en effet calculé à partir des résultats d'un échantillon de la population. Il se peut qu'au sein de cette population la corrélation entre les variables soit nulle et que le coefficient observé soit différent de zéro du seul fait du hasard. Il est donc nécessaire de tester l'hypothèse selon laquelle la corrélation est effectivement nulle. Pour ce faire, on peut estimer la valeur de t à l'aide de la formule suivante :

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2.25)$$

Cette valeur se distribue comme t avec $n-2$ degrés de liberté. Nous pouvons dès lors comparer la valeur obtenue avec les valeurs critiques de la distribution t de Student pour le nombre de degrés de liberté adéquat. Si la valeur obtenue est supérieure à la valeur critique, nous pourrons considérer que le coefficient de corrélation observé est significativement différent de zéro. Appliquons cette formule au cas du coefficient de corrélation calculé ci-dessus :

$$t = \frac{0{,}540\sqrt{92-2}}{\sqrt{1-(0{,}540)^2}} = 7{,}237 \qquad (2.26)$$

Cette valeur est significative au seuil de 0,001. Par conséquent, nous pouvons affirmer qu'il existe bien une relation non nulle entre les performances au test de calcul mental et à celui de mémoire de chiffres pour les enfants de 8 ans et demi.

Mais qu'une corrélation soit significative n'implique pas qu'il existe une relation étroite entre les variables considérées. Pour interpréter correctement la relation entre les variables, il est utile de calculer le *coefficient de détermination (r^2)* qui est égal au carré du coefficient de corrélation. La valeur r^2peut en effet être interprétée comme la proportion de variance d'une des mesures qui est liée à la variance de l'autre mesure. Par exemple, la corrélation de 0,540 entre les deux tests présentés plus haut signifie que 25% (c'est-à-dire $0{,}540^2$) de la variance des scores à l'un des tests est liée à la variance des scores à l'autre test. Par conséquent, 75% de la variance observée sur la première variable est sans relation linéaire avec la seconde variable. Pour illustrer d'une autre manière la même idée, nous pouvons dire que, connaissant les résultats au test de calcul mental, nous ne pouvons prédire que 25% de la variance des scores au test de mémoire de chiffres (et réciproquement). Cette manière d'aborder les coefficients de corrélation nous permet d'avoir une idée plus juste de leur importance. Souvent des coefficients sont significatifs mais ne nous apportent que peu d'information. Par exemple, une coefficient de 0,25 signifie que seulement 6% de la variance est partagée par les deux variables considérées.

Parlant des corrélations entre variables, nous avons utilisé des termes comme " liaison ", " association ", " prédiction " en évitant soigneusement d'inférer une relation de cause à effet entre les variables. En fait, l'explication de la relation observée entre deux variables est une question extérieure à la statistique. Cette interprétation doit se faire sur base d'un modèle théorique de la réalité étudiée. Dans l'exemple ci-dessus, nous pourrions interpréter la corrélation observée en nous appuyant sur un modèle théorique de la résolution de problèmes arithmétiques. Dans certains cas, nous pourrons avancer l'hypothèse d'une relation de cause à effet entre les variables. Mais, souvent, nous devrons postuler le rôle de variables sous-jacentes aux variables observées pour expliquer la liaison entre celles-ci. Par exemple, nous pourrions expliquer la corrélation entre les scores à un test d'arithmétique et à un test de langue maternelle par la variable " année d'étude " ou par la variable " intelligence " (ou encore par une interaction de ces deux variables). Parfois, certaines corrélations ne sont pas interprétables car elles sont le fruit du seul hasard. Par exemple, en Allemagne, après-guerre, on a observé une relation entre le nombre de cigognes et le nombre de naissances. Dans ce cas, aucune théorie sérieuse ne permettait d'expliquer cette association purement fortuite.

Dans certains cas, le coefficient de corrélation peut être sous-évalué du fait de la *réduction de l'étendue des scores* sur l'une des variables. En psychométrie, nous avons affaire à une réduction de l'étendue lorsque les résultats d'un groupe particulier se concentrent sur une zone étroite de l'étendue possible des scores. Cette situation se présente fréquemment lorsque l'on veut valider des tests de sélection en entreprise ou en éducation. Par exemple, il est logique de vouloir évaluer la validité prédictive d'un exa-

men d'entrée dans l'enseignement supérieur en calculant la corrélation entre les scores à cet examen et la moyenne de résultats en fin de première année. Toutefois, en procédant de la sorte, on sous-évalue automatiquement la corrélation entre les deux variables concernées. En effet, seuls les meilleurs étudiants ont été sélectionnés sur base de leurs résultats à l'examen d'entrée. Par conséquent, les résultats des examens de fin de première année présentent une variabilité sensiblement réduite puisque les étudiants les plus faibles n'ont pas eu l'opportunité de passer ces examens.

Une illustration graphique permet de comprendre aisément pourquoi la réduction de l'étendue des scores entraîne une sous-estimation du coefficient de corrélation. La figure 6 présente le diagramme cartésien pour deux séries de scores obtenus par un échantillon de sujets. Lorsque nous observons le nuage de points pour l'ensemble du groupe, nous remarquons la forme elliptique caractéristique d'une liaison positive d'intensité moyenne entre les deux variables (le coefficient de corrélation est ici égal à 0,60). Si, à présent, nous ne nous intéressons qu'aux sujets se situant dans le tiers supérieur de la distribution des scores de la variable X (partie encadrée), le nuage de point n'est plus du tout elliptique ce qui indique une très faible corrélation entre les deux variables.

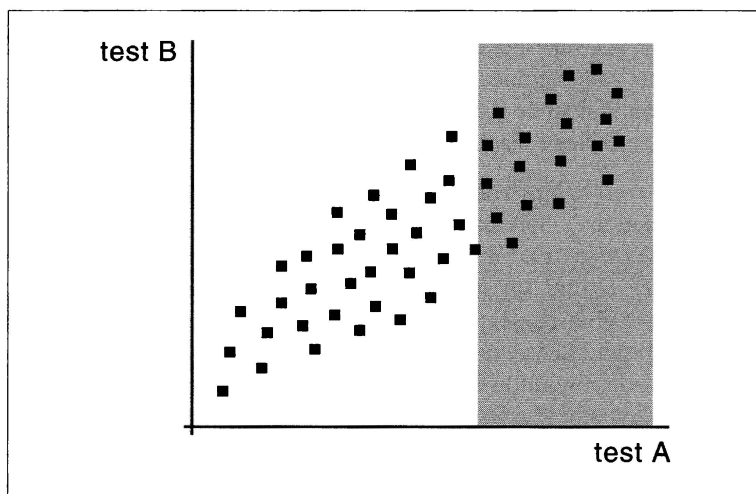


Figure 6 – Effet de la réduction de l'étendue des scores sur le coefficient de corrélation

Dans certains cas, le coefficient de corrélation peut chuter dramatiquement lorsque l'étendue des scores est fortement réduite. Un exemple célèbre est donné par Torndike (1949, pp. 170-171) concernant un programme de sélection de l'US Air Force. Une batterie de tests avait été constituée pour prédire le succès dans l'apprentissage du pilotage. Sur base des résultats à ces tests, seuls 13% des candidats étaient suffisamment qualifiés pour être admis dans le programme d'apprentissage. Toutefois, dans un but expérimental, on décida d'admettre tous les candidats. A la fin de la période d'entraînement, on évalua les qualités de pilote de chacun et l'on calcula les corrélations entre ce critère et les résultats aux différents tests. Ces corrélations furent calculées pour l'ensemble du groupe (N=1036) et pour le groupe des meilleurs candidats (N=136). On constata ainsi que la corrélation entre le critère et le test de coordination

complexe était de 0,40 pour l'ensemble du groupe et de -0,03 pour le groupe restreint. De même, la corrélation entre le critère et le score composite d'aptitude était de 0,68 pour l'ensemble du groupe et de seulement 0,18 pour le groupe des meilleurs candidats. La valeur des prédictions réalisées à l'aide de la batterie de tests était donc très faible si l'on se basait sur les seuls résultats des candidats les plus brillants. Par contre, cette même qualité des prédictions était satisfaisante lorsque l'on évitait la réduction de l'étendue des scores en calculant les coefficients de corrélation à partir des résultats de l'ensemble du groupe.

Dans l'exemple que nous venons de citer, il a été possible d'évaluer la corrélation correcte entre les variables puisque les chercheurs possédaient les résultats pour l'ensemble du groupe. Malheureusement, cette information fait souvent défaut dans les études de validité. C'est, par exemple, ce qui se passe pour les tests d'admission. Dans ce cas, nous ne possédons les résultats au critère que pour les sujets qui ont été sélectionnés sur base du test initial. Il est toutefois possible de corriger le coefficient obtenu sur l'échantillon restreint et d'obtenir une meilleure estimation de la validité du test. Le coefficient corrigé n'est cependant qu'une approximation et doit être utilisé avec prudence.

4.3 LA DROITE DE RÉGRESSION

Lorsque la relation entre deux variables est assez étroite et linéaire, il est intéressant de représenter cette relation sous la forme d'une droite traversant le nuage de points. Cette ligne, appelée *droite de régression*, est la meilleure ligne droite représentant les différentes coordonnées du diagramme cartésien. La figure 7 présente la relation entre les résultats d'un échantillon de 100 sujets âgés de 65 à 69 ans aux épreuves d'information et de vocabulaire du test d'intelligence WAIS-R. La corrélation entre ces deux variables est égale à 0,869. Au sein du nuage de points est tracée la droite de régression.

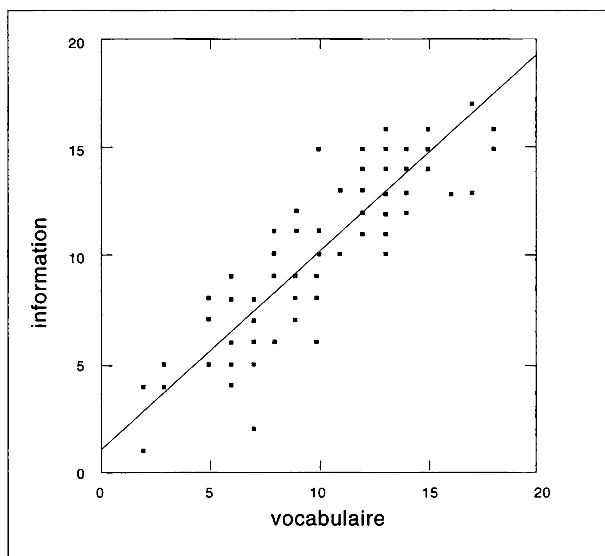


Figure 7 – Exemple de droite de régression

La droite de régression est très utile lorsque nous souhaitons prédire les résultats sur une variable à partir des scores sur l'autre variable. Cette technique est souvent utilisée avec les tests de sélection et d'orientation. Par exemple, sur base des performances à un test de mathématique, on peut estimer les futurs résultats dans une section scientifique. Cette prédiction constitue une information intéressante pour aider les étudiants à s'orienter dans leurs études.

La droite de régression est définie par une équation de la forme $Y = bX + a$. Dans le cas présent, cette équation s'écrit :

$$\hat{Y} = bX + a \tag{2.27}$$

$\hat{Y}$ = la valeur de Y prédite à partir de X (l'accent circonflexe sur Y indique que les valeurs de Y calculées à partir de X ne sont que des estimations des valeurs exactes de Y).

b = la pente de la droite de régression (elle correspond à la différence de valeur sur l'ordonnée associée à une différence d'une unité sur l'abscisse).

a = l'intersection de la droite avec l'ordonnée (elle est égale à la valeur de Y lorsque $X = 0$).

Pour déterminer la droite de régression la plus proche possible des valeurs effectives de Y, il nous faut trouver les valeurs de a et de b qui définissent la fonction linéaire la plus adéquate. En d'autres termes, nous devons déterminer les valeurs a et b qui minimisent l'erreur de prédiction de Y à partir de X. Cette erreur peut être évaluée à partir de la formule suivante :

$$\text{erreur de prédiction} = \sum (Y - \hat{Y})^2 \tag{2.28}$$

Cette quantité, permettant de sélectionner la meilleure fonction linéaire, est appelé *le critère des moindres carrés*. Les valeurs de a et de b qui minimisent cette quantité peuvent être trouvées au moyen des formules suivantes :

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \tag{2.29}$$

$$a = \frac{\sum Y - b\sum X}{n} = \bar{Y} - b\bar{X} \tag{2.30}$$

Appliquons ces formules à l'exemple ci-dessus, dont un extrait des données et le résultat de quelques calculs intermédiaires sont présentés dans le tableau 9 :

$$b = \frac{(100 \times 12106) - (1035 \times 1051)}{(100 \times 12081) - (1035)^2} = 0,897 \tag{2.31}$$

$$a = \frac{1051 - (0,897 \times 1035)}{100} = 1,226 \tag{2.32}$$

Dans ce cas, l'équation de régression peut s'écrire : $\hat{Y} = (0,897)X + 1,226$

Grâce à cette équation, nous pouvons maintenant estimer les valeurs de Y pour chaque valeur de X. Par exemple, si $X = 2$ alors $\hat{Y} = 3,02$ et si $X = 14$ alors $\hat{Y} = 12,558$.

Il est important de souligner que l'équation de régression que nous avons déterminée ci-dessus nous permet d'estimer les valeurs de Y à partir des valeurs de X mais non l'inverse. Si nous voulons estimer X à partir de Y, il nous faut estimer les paramètres qui minimisent $\sum (X - \hat{X})^2$. Par conséquent, les droites de régression de Y sur X et de X sur Y ne coïncident habituellement pas.

Tableau 9 – Extrait des résultats aux tests de vocabulaire (X) et d'information (Y) (N=100)

Sujets	Test de vocabulaire	Test d'information
1	15	12
2	11	8
3	8	7
4	6	7
5	7	5
6	16	15
7	7	7
8	13	16
9	12	14
10	15	13
...	...	...

$$\sum X = 1035 \qquad \sum X^2 = 12081 \qquad \sum Y = 1051 \qquad \sum XY = 12106 \qquad (2.34)$$

Par ailleurs, nous ne devons pas oublier que les valeurs de Y que nous calculons à l'aide de l'équation de régression ne sont que des estimations des valeurs réelles. Les coordonnées des valeurs de X et des estimations de Y forment une droite parfaite alors que les valeurs effectives de Y se dispersent autour de cette droite. En fait, les valeurs que nous obtiendrions si nous pouvions mesurer directement la variable Y se distribuent normalement autour des valeurs estimées. Les distributions de Y autour de chaque valeur estimée sont appelées les *distributions conditionnelles* de Y. La figure 8 permet de mieux comprendre ce que représentent ces distributions. Pour trois estimations de Y, nous avons tracé la distribution de fréquences des valeurs effectives de Y. Nous pouvons constater que la moyenne de ces distributions correspond à la valeur estimée. Quant à l'écart-type de ces distributions, il nous informe sur l'erreur de notre estimation. Plus cet écart-type est important, plus notre estimation risque d'être éloignée de la valeur que nous aurions pu obtenir en mesurant directement Y.

Cette erreur d'estimation est très utile pour le praticien. À l'aide de cette erreur, celui-ci peut construire un intervalle de confiance autour de la valeur estimée. Il peut ainsi se faire idée de l'approximation de son estimation de Y à partir de X. Ceci est important lorsque nous utilisons les résultats d'un test dans un but de prédiction. L'usage systématique de l'intervalle de confiance nous conduit en effet à une plus grande prudence dans nos décisions. L'erreur type d'estimation peut être calculée à l'aide de la formule suivante :

$$s_{Y \cdot X} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n - 2}} \tag{2.34}$$

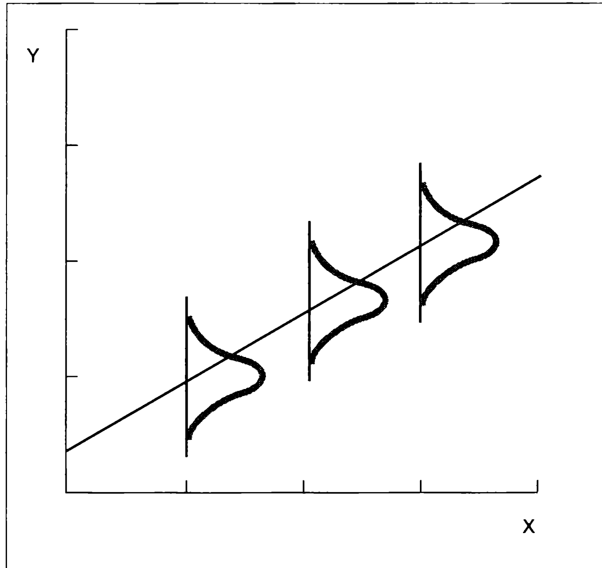


Figure 8 – Exemples de distributions conditionnelles de Y

Dans notre exemple, l'erreur-type d'estimation est égale à 1,909. Cela signifie que si, pour $X = 2$, la valeur estimée de Y est égale à 3,017, nous pouvons en déduire qu'environ 68% des valeurs effectives de Y sont incluses dans l'intervalle compris entre (3,017 - 1,909) et (3,017 + 1,909). Rappelons en effet que, si la distribution est normale, 68% des valeurs observées se situent dans l'intervalle de moins un écart-type et plus un écart-type autour de la moyenne. Si nous souhaitons un intervalle incluant 95% des valeurs autour de la moyenne, il nous suffit de prendre deux écarts-types autour de la valeur observée. Dans notre exemple, les bornes seront dès lors : (3,017 - 2(1,909)) et (3,017 + 2(1,909)), c'est-à-dire 0,807 et 6,835. Concrètement, cela signifie que, sur base d'un résultat égal à 2 au test de vocabulaire, nous pouvons prédire que le résultat au test d'information sera égal à 3. Mais les résultats que nous pourrons effectivement observer à ce test auront 95% de chances de se situer entre 1 et 7 points. Une telle observation doit nous inciter à la prudence lorsque nous utilisons une valeur estimée pour prendre des décisions.

L'usage d'un intervalle de confiance unique, quelle que soit la valeur estimée, repose sur deux postulats : (1) les distributions conditionnelles de Y sont normalement distribuées, (2) les variances de toutes ces distributions sont égales. Ce dernier postulat d'homogénéité de la variance (appelé aussi *postulat d'homoscédasticité*) est souvent difficile à satisfaire parfaitement avec des données réelles. Par conséquent, l'usage d'un intervalle de confiance unique peut conduire à des erreurs. Certains auteurs recommandent par conséquent de calculer l'erreur de mesure pour chaque estimation de Y (voir par exemple Howell, 1992, p. 244-245, pour une description de ce calcul).

Cette procédure, trop lourde pour un usage routinier, peut être recommandée lorsque l'on définit des scores " seuils " dans un test de sélection ou d'admission.

5. Le choix de la bonne méthode statistique

La grande variété des techniques statistiques disponibles rend complexe le choix de la méthode qui convient le mieux à un test d'hypothèse. Ce choix est d'autant plus important qu'il influence directement le type d'erreur et la puissance de nos décisions statistiques. Quoique ce chapitre et le précédent n'ont fait qu'introduire un petit nombre de techniques statistiques parmi les plus répandues et les plus souvent employées, il est important de bien les situer dans un contexte global ainsi que les unes ar rapport aux autres.

Tableau 10 – Synthèse des principales techniques statistiques

	Statistique univariée *(1 variable dépendante)*	**Statistique univariée** *(1 variable dépendante, 1 variable indépendante ou plus)*	**Statistique mutivariée** *(plusieurs variables dépendantes)*
Statistiques descriptives	Moyenne, variance, écart-type	Corrélation simple (de Pearson) Corrélation par rangs de Spearman	Analyse factorielle exploratoire
Statistiques inférentielles paramétriques	Estimation de la moyenne Estimation de la variance	Test t, ANOVA Test de signification sur la corrélation Analyse de covariance, analyse de régression	Analyse factorielle confirmatoire Analyse de variance multivariée
Statistiques inférentielles non paramétriques	Test du χ^2 pour un échantillon	Test du χ^2 pour deux échantillons Test de signification du W de Kendall, rho de Spearman	Analyse de correspondance (dual scaling)

Le tableau 10 propose d'organiser les principales techniques statistiques sous forme d'un tableau à double entrée. Les rangées distinguent entre *statistiques descriptives* et *statistiques inférentielles*. Les colonnes identifient les variables prises en considération par chacune des techniques.

Parmi les statistiques descriptives, nous retrouvons toutes les valeurs de distribution pour une seule variable dépendante que nous avons étudiées dans le chapitre 1 : moyenne, écart-type, variance. Nous y retrouvons aussi la corrélation entre deux variables et un prolongement de cette technique à plusieurs variables dépendantes, l'analyse factorielle exploratoire. L'analyse factorielle exploratoire permet d'identifier à partir d'une *matrice de corrélations* les *traits latents* qui permettent de regrouper plusieurs variables dépendantes en un petit nombre de *facteurs* indépendants. Il en sera question au chapitre 5.

Bien souvent, cependant, nous sommes intéressés à aller au-delà de la description d'un échantillon. Nous voulons déduire certaines caractéristiques de la population à partir de celles de l'échantillon. C'est l'objet de l'ensemble des statistiques inférentielles, dont ce chapitre a été l'objet. Parmi les techniques impliquant une seule variable dépendante, nous retrouvons l'estimation de la moyenne de la population. Parmi les techniques impliquant une variable dépendante et une variable indépendante, on retrouve les tests de comparaison de moyennes (tests *t* de Student) et l'analyse de variance (ANOVA) qui peut impliquer plus d'une variable indépendante. On retrouve également dans cette catégorie le test de signification d'une corrélation et toutes les techniques apparentées à l'*analyse de régression*, dont la *régression logistique* des *modèles de réponses aux items* (chapitre 8) constitue un cas particulier. Il existe aussi toute la famille des *statistiques multivariées* dont nous ne ferons pas état dans ce livre. Cette famille regroupe toutes les techniques statistiques permettant le test d'hypothèses portant sur plusieurs variables dépendantes : c'est le cas de l'*analyse de variance multivariée* et de l'*analyse discriminante*.

Enfin, nous pourrions établir une distinction supplémentaire à l'intérieur de la catégorie des statistiques inférentielles. On peut différencier ces techniques statistiques selon qu'elles font appel à l'estimation des paramètres de la population ou non. Dans le premier cas, nous parlerons de *statistiques paramétriques* : c'est le cas de toutes les techniques que nous avons vues jusqu'à présent. Toutes reposent sur des échantillons dont les résultats se distribuent normalement. Toutes font appel au calcul des principaux paramètres de la distribution normale que sont la moyenne et la variance. Dans le second cas, nous parlerons de *statistiques non paramétriques*. Cette catégorie d'outils statistiques permet le test d'hypothèses en l'absence de postulats concernant la distribution de la population et ses principaux paramètres. C'est le cas des tests de comparaison de fréquences (*test du* χ^2 *- khi-carré-)* ou des médianes. C'est le cas aussi des coefficients de corrélation par rangs tels que le *rho de Spearman* (chapitre 6) ou le *W de Kendall* (chapitre 5). Il existe plusieurs ouvrages discutant des propriétés de ces outils statistiques, particulièrement puissants avec des échantillons restreints ($n < 30$). Pour une bonne introduction à l'ensemble de ces outils statistiques, nous recommandons l'ouvrage de Siegel et Castellan (1988).

Plusieurs nouvelles techniques statistiques seront expliquées dans les prochains chapitres. Les théories de la mesure font appel à l'application de ces techniques à des problèmes particuliers de quantification. Avec la conclusion du chapitre 2, nous postulons que les principaux éléments de *statistique théorique* sont assimilés. Nous avons restreint au minimum les aspects théoriques afin de pouvoir nous concentrer sur les problèmes de *statistique appliquée* que posent la mesure en psychologie et en éducation. Le lecteur qui souhaite approfondir les fondements théoriques des techniques abordées pourra faire appel aux nombreuses références à la fin de ce livre.

CHAPITRE 3

LA CONSTRUCTION D'UN INSTRUMENT DE MESURE

1. Le processus de construction d'un test

La construction d'un test en psychologie ou en éducation est un processus de longue haleine. Cinq étapes principales peuvent être distinguées dans ce processus. Cette section se limite à une brève présentation de chacune de ces étapes. Les premières étapes seront analysées plus en détail dans les sections suivantes du présent chapitre. Les autres feront l'objet des chapitres 4 à 7.

Étape 1 : La détermination des utilisations prévues du test

La première question que doit se poser la personne désireuse de construire un test concerne les fonctions que ce dernier devra remplir. A quoi va-t-il servir ? Par exemple, un test de mathématique peut avoir pour fonction de sélectionner des sujets, de diagnostiquer des difficultés d'apprentissage ou encore de déterminer si un élève maîtrise les compétences attendues en fin d'année scolaire. De même, un questionnaire d'anxiété peut être utilisé pour recruter des personnes possédant certaines caractéristiques de personnalité ou pour évaluer l'effet d'un médicament anxiolytique. Le plus souvent, un même test ne peut remplir toutes ces fonctions. En effet, les usages prévisibles d'un test déterminent profondément ses caractéristiques. En particulier, un distinction nette doit être tracée entre les tests normés et les tests critériés. Les *tests normés* visent à discriminer les sujets appartenant à la population pour laquelle est construite le test. Ces tests peuvent, par exemple, nous procurer des informations sur le

degré d'anxiété d'un sujet par rapport au niveau de l'anxiété dans l'ensemble de la population. Il en va de même pour le niveau de compétence en mathématique ou pour tout autre caractéristique que l'on souhaite mesurer. Par contre, les *tests critériés* ont pour fonction d'évaluer si un sujet possède ou non certaines caractéristiques prises comme référence. Par exemple, pour remplir correctement une certaine fonction professionnelle, le niveau d'anxiété du sujet ne dépasse-t-il pas un seuil déterminé ? Ou encore, le sujet possède-t-il les compétences en mathématique nécessaires pour aborder un programme d'étude donné ?... Le choix de construire un test normé ou un test critérié conditionne la méthodologie utilisée. Des techniques particulières doivent être appliquées pour obtenir des tests possédant des propriétés métriques spécifiques.

La distinction entre test normé et test critérié n'est pas la seule qui puisse être faite. Dans le domaine éducatif, il existe de profondes différences entre les tests destinés à l'évaluation certificative et ceux utilisés pour l'évaluation formative ou l'évaluation diagnostique. Un *test certificatif* doit couvrir l'ensemble d'un programme scolaire. Un tel test est habituellement centré sur les performances. Il doit en effet permettre de vérifier si l'élève est capable de réaliser les tâches que l'on attend de lui en fin d'apprentissage. Par contre, un *test diagnostique* est généralement beaucoup plus ciblé. Son but est de comprendre le sens d'une performance. Par exemple, il ne s'agit plus, comme avec un test certificatif, de simplement vérifier si un élève peut additionner correctement deux nombres décimaux, mais de comprendre pourquoi certains élèves présentent des difficultés pour réaliser de telles additions. L'information que l'on désire recueillir ne se limite plus à la performance mais concerne les capacités cognitives sous-jacentes à ces performances. Pour atteindre cet objectif, il est nécessaire d'utiliser un test qui s'appuie sur un modèle des processus mis en jeu pour réaliser des additions avec des décimaux. Un tel modèle permet d'éclairer les difficultés rencontrées par les élèves et, le cas échéant, de mettre en oeuvre des actions remédiatives. Ainsi, les propriétés d'un test diagnostique sont nécessairement très différentes de celles-ci d'un test certificatif. Ces deux types d'outils doivent, par conséquent, être conçus de manière spécifique en s'appuyant sur une méthodologie adaptée.

Il est possible d'opérer d'autres distinctions entre les fonctions que peuvent remplir les tests. Comme nous venons de le voir, ces fonctions déterminent la nature du test à construire et, par conséquent, la méthodologie à utiliser pour élaborer un tel outil. On ne peut donc éluder une réflexion approfondie sur l'usage auquel on destine un test. Au point de départ du travail de construction, un choix doit toujours être opéré entre différentes fonctions possibles. Il est illusoire de vouloir créer un test « généraliste » qui ambitionne de répondre à tous les besoins des praticiens. Dans la section 2 du présent chapitre, cette question sera approfondie dans le cas du développement d'un test d'acquis scolaires.

Étape 2 : La définition de ce que l'on souhaite mesurer

Habituellement, le point de départ d'un test est un objectif relativement vague et général : « évaluer la compréhension en lecture à l'école primaire », « apprécier le développement social de 3 à 6 ans », « diagnostiquer les troubles de la mémoire », « sélectionner des secrétaires »... Ces intentions sont encore beaucoup trop vagues pour permettre réellement de débuter la construction d'un test. Elles nécessitent un tra-

vail d'approfondissement des concepts et d'opérationnalisation de ceux-ci. En d'autres termes, il s'agit de définir avec précisions les caractéristiques psychologiques ou éducatives que le test devra mesurer. Sur base de cette définition, des items pourront alors être construits. Cette première étape est donc cruciale. Nous verrons dans le chapitre 5 que la validation du contenu du test repose sur ce travail préalable de définition de ce que l'on veut mesurer.

Mais comment passer d'une intention vague à la définition opérationnelle d'un concept ? Selon les domaines, plusieurs méthodes peuvent être utilisées :

1. *La définition des objectifs pédagogiques et la construction d'un tableau de spécifications.* Lorsqu'il s'agit d'évaluer des apprentissages scolaires, la démarche la plus fréquente consiste à préciser les performances dont les élèves devront être capables à un moment donné de leur apprentissage. De nombreux outils ont été développés pour permettre une opérationnalisation suffisante de ces objectifs. Le tableau de spécifications est un de ces outils permettant de déterminer les divers type de comportements attendus relativement à un contenu disciplinaire. La section 2 du présent chapitre présente en détail la construction d'un tableau de spécification ainsi que d'autres méthodes permettant de préciser les caractéristiques que doit évaluer un test d'acquis scolaire.

2. *L'analyse de contenu d'entretiens.* Lorsque le praticien n'a pas d'idées précises à propos des caractéristiques permettant de discriminer les sujets qui seront évalués par le test, il est intéressant de commencer par interroger des personnes appartenant à la population visée par ce test. L'interview, libre ou semi-structurée, permet de recueillir un grand nombre d'informations qui seront sélectionnées et classées au moyen d'une analyse de contenu. Par exemple, Hunt & McKenna (1992) ont procédé de la sorte pour mettre au point un questionnaire de qualité de vie destiné à des patients dépressifs. Cinq psychiatres ont interviewé 30 patients dépressifs à propos de différentes facettes de leur vie quotidienne. Une analyse de contenu des entretiens a permis de mettre en évidence un certain nombre de propositions caractéristiques, permettant d'apprécier la qualité de vie des patients dépressifs. Ces propositions ont ensuite servi à construire les items du questionnaire.

3. *L'observation directe des comportements.* Dans certains cas, plutôt que d'interroger les sujets, il est préférable de les observer dans leur milieu de vie ou de travail. Cette méthode a été utilisée par Binet pour construire le tout premier test d'intelligence de l'histoire. Au début de ce siècle, Binet ne pouvait s'appuyer que sur un modèle rudimentaire et vague de l'intelligence. Dès 1900, il commença donc à observer les arriérés adultes de l'Asile Sainte-Anne et les enfants d'une école d'un quartier populaire Paris afin de mettre en évidence les comportements permettant de distinguer les sujets normaux des sujets handicapés mentaux. Les items de l'échelle métrique d'intelligence de 1905 sont issus de ce travail d'observation.

4. *La méthode des incidents critiques.* L'origine de cette méthode est attribuée à Flanagan (1954). Elle est particulièrement utile pour construire des outils d'évaluation des performances professionnelles. Elle consiste à demander à des

responsables de décrire des situations de travail où les personnes sous leurs ordres ont agi de manière particulièrement efficace ou, au contraire, inefficace. Partant de cette description, certains comportements « critiques » peuvent être mis en évidence et servir pour construire des échelles d'évaluation.

5. *La référence à un modèle théorique.* À la différence des autres méthodes celle-ci ne part pas de l'expérience mais d'un modèle de la réalité construit au cours de recherches antérieures. Depuis le début des années 80, les développements de la psychologie cognitive ont conduit à la création de nombreux modèles théoriques utilisables par les constructeurs de tests. Des tests destinés au diagnostic des troubles de la lecture ont, par exemple, été créés sur base de modèles décrivant les processus impliqués dans l'activité de lecture (p.e. : de Partz, 1994 ; Mousty & al. 1994...). D'autres outils ont également été construits en référence à des modèles théoriques pour évaluer des caractéristiques aussi diverses que le calcul, la motivation, la mémoire...

Étape 3 : La création des items

Il y a près de cinquante ans, Georges Gallup, fondateur du célèbre institut de sondage du même nom, affirmait (1947, p.383) : « *Trop d'attention a été accordée à la constitution des échantillons et trop peu à la création des questions [...]. Des différences dans la construction des questions conduisent souvent à des résultats qui présentent de plus grandes variations que celles habituellement observées en fonction des différentes techniques d'échantillonnage* ». Cette constatation garde toute son actualité et peut être généralisée aux questions construites pour les tests psychologiques et les tests d'acquis scolaires. Souvent, les praticiens ne suivent aucune méthodologie pour construire les items. Ayant en tête ce qu'ils souhaitent mesurer, ils se fient à leur intuition pour produire les questions. Pourtant, il est indispensable d'avoir les idées claires sur plusieurs points avant de se lancer dans la production d'items :

1. *Quel format d'items choisir ? Pourquoi ?* Le choix d'un format ne doit pas être arbitraire. Il découle d'un ensemble de contraintes concernant les objectifs du test et les conditions matérielles de création, de passation et de cotation de celui-ci. En conséquence, il n'y a pas de bon format d'item dans l'absolu. Un format est bon s'il est adéquat au but et à la situation d'évaluation. La section 3 du présent chapitre aborde de manière détaillée la question du choix du format d'item et des règles de construction de différents formats de questions fermées et ouvertes.

2. *Quel doit être le niveau de difficulté des items ?* Le choix du niveau de difficulté des items dépend de l'objectif du test. Ce niveau variera selon que le test est normé ou critérié, certificatif ou formatif... En d'autres termes, c'est la nature des informations que l'on désire recueillir qui doit déterminer le niveau de difficulté des items à produire.

3. *Combien faut-il créer d'items ?* Le nombre d'items à créer dépend de plusieurs facteurs. Le premier facteur est la durée du test. Selon que l'on souhaite un test court, pouvant être passé en 10 minutes, ou un test diagnostic se déroulant sur plusieurs séances d'examen, le nombre d'items à créer variera considérablement. Un second facteur à prendre en compte est le niveau de fiabilité du test

désiré. Un test long sera généralement plus fiable qu'un test court (voir chapitre 4). Par ailleurs, si le test comporte plusieurs sous-scores, il sera nécessaire d'assurer la fiabilité de ceux-ci en prévoyant suffisamment d'items dans chacune des sous-échelles du test. Enfin, un dernier facteur à prendre en considération est l'élimination, quasi inévitable, de certains items après leur évaluation par des experts et le prétest. Si l'on veut que la version finale du test contienne assez d'items, il faudra donc en créer plus que le strict nécessaire. Si, par exemple, le test final doit contenir 20 items, on en créera 30 et l'on retiendra les 20 meilleurs de ceux-ci. Habituellement, un surplus de 30 à 50% d'items est nécessaire pour éviter les mauvaises surprises après le prétest.

Étape 4 : L'évaluation des items

Une définition précise de ce que l'on souhaite mesurer et une méthodologie rigoureuse de construction des items sont des conditions nécessaires mais non suffisantes pour obtenir des items valides et fiables. Pour garantir les propriétés métriques des items, une évaluation minutieuse de ceux-ci doit également être réalisée. Deux démarches complémentaires sont habituellement suivies pour réaliser cette tâche :

1. *Une évaluation des items par des juges.* Ceux-ci sont chargés d'apprécier la conformité des items aux exigences définies lors de la seconde étape du processus de construction du test. Les méthodes d'évaluation des items par des juges sont détaillées dans la section 2 du chapitre 5 consacré à la validité.
2. *La réalisation d'un prétest des items* suivie d'une analyse qualitative et quantitative des résultats. Le prétest complète l'appréciation des items par des juges. Cette dernière évaluation reste en effet subjective malgré la rigueur méthodologique avec laquelle elle peut être réalisée. Le prétest permet, lui, de recueillir des données empiriques, directement au sein de la population à laquelle est destinée le test.

Le prétest consiste à faire passer tous les items à un échantillon de la population. Cet échantillon ne doit pas nécessairement être représentatif (voir chapitre 7 pour une discussion de cette notion) ni de très grande taille. La taille de cet échantillon dépend en fait de l'hétérogénéité de la population visée par le test. Par exemple, si un questionnaire de stress est destiné à évaluer uniquement des pilotes d'avion, un prétest sur un échantillon de 50 sujets permettra généralement une évaluation satisfaisante des items. Par contre, si la population est plus hétérogène, un échantillon de 200 à 300 sujets peut être nécessaire pour réaliser un prétest valable. Par exemple, le prétest des items de la version française du WISC-III (Wechsler Intelligence Scale for Children - version 3) a été réalisé sur un échantillon de 220 sujets. Ce test est destiné à évaluer tous les enfants français entre 6 et 16 ans. Dans ce cas, l'échantillon du prétest doit être de plus grande taille car il doit inclure des enfants des deux sexes, de différents âges et de différents milieux sociaux. On ne vise toutefois pas à ce qu'un tel échantillon soit parfaitement représentatif de la population. Il doit avant tout refléter l'hétérogénéité de celle-ci. Un échantillon trop homogène risque en effet de masquer certains items problématiques. Par exemple, si les items d'un questionnaire de dépression destiné à des personnes âgées sont prétestés sur un échantillon qui ne comprend que des retraités ayant terminés des études supé-

rieures, certains problèmes risquent de passer inaperçus. L'inclusion de personnes âgées possédant le seul diplôme d'études primaires aurait permis de mettre en évidence des questions dont le vocabulaire trop complexe peut entraîner des erreurs de compréhension.

Les résultats du prétest sont analysés d'un point de vue tant qualitatif que quantitatif. En particulier, les commentaires des sujets à propos des items peuvent se révéler précieux pour comprendre des résultats aberrants et pour remédier à certains problèmes de formulation des questions. De même, les problèmes de manipulation du matériel, d'enregistrement des réponses, de temps de passation, de cotation des réponses... peuvent être repérés à l'occasion du prétest. Ces problèmes, en apparence mineurs, doivent retenir toute l'attention du constructeur car ils peuvent diminuer considérablement la validité des résultats d'un test. C'est, par exemple, le cas d'un espace trop petit pour noter les réponses ou d'un livret de test difficile à manipuler.

En plus de ces vérifications qualitatives, le prétest permet de réaliser différentes analyses statistiques des résultats. Celles-ci sont détaillées dans le chapitre 6 consacré à l'analyse des items. Ces analyses portent sur la difficulté des items, leur discrimination, leur fonctionnement différentiel.... Sur base de ces analyses et des observations qualitatives, les meilleurs items seront finalement sélectionnés et serviront à construire la version définitive du test.

Étape 5 : La détermination des propriétés métriques du test définitif

Une fois les meilleurs items sélectionnés et la version définitive du test constituée, il reste encore à déterminer les propriétés métriques de ce test. Les propriétés qui doivent retenir l'attention du constructeur varient en fonction de la nature du test. S'il s'agit d'un test normé, il sera nécessaire d'établir des normes et de présenter celles-ci selon une échelle aisément compréhensible par les praticiens. S'il s'agit d'un test critérié, il faudra préciser des scores de référence utiles aux praticiens. Par ailleurs, si les résultats du test doivent être mis en relation avec ceux d'autres tests, il y aura lieu de mettre en équivalence les échelles de mesure concernées. Les techniques nécessaires pour déterminer les normes, les scores de référence et les équivalences sont présentées en détail dans le chapitre 7.

Par ailleurs, une investigation approfondie de la validité et de la fiabilité de la version finale du test devra toujours être réalisée. Les études de validité concerneront la validité en référence à un critère externe et la validité conceptuelle. Le constructeur doit en effet rassembler des preuves de la validité des inférences qu'il suggère de réaliser à partir des résultats au test. Par exemple, s'il propose aux praticiens de calculer et d'interpréter différents sous-scores au test, il sera nécessaire de prouver la pertinence de tels sous-scores et de la signification qui leur est attribuée (American Psychological Association, 1985, p.14). Les fondements et la méthodologie de telles études de validité sont explicités dans les sections 3 et 4 du chapitre 5. Il faut de souligner que l'évaluation de la validité d'un test n'est pas du seul ressort du constructeur. En fait, la validité n'est jamais une qualité acquise une fois pour toute. Chaque nouvelle inférence qu'un praticien veut réaliser à partir des résultats d'un test doit faire l'objet d'une validation spécifique. Par exemple, si un test de mémoire a été créé pour évaluer les

compétences mnésiques des enfants et des adolescents, la pertinence de l'usage de ce test avec des adultes devra être argumentée sur base de données empiriques.

Le constructeur devra également apporter des informations à propos de la fiabilité du test. Il ne s'agit pas uniquement d'un coefficient de fiabilité mais aussi de mesures liées à celui-ci et nécessaires aux praticiens, comme l'erreur de mesure de scores, les intervalles de confiances, l'erreur de mesure des différences entre scores... Les techniques nécessaires pour calculer ces valeurs relatives à la fiabilité sont présentées de manière détaillée dans le chapitre 4.

Lorsqu'un test n'est pas réservé au seul usage de son constructeur mais est destiné à être diffusé, la rédaction d'un manuel est nécessaire (American Psychological Association, 1985, pp.35-37). Ce manuel doit non seulement présenter les données métriques citées ci-dessus (normes, coefficient de fiabilité...) mais doit également détailler les bases théoriques du test, les fonctions pour lesquelles il a été créé et les qualifications requises pour pouvoir l'appliquer et l'interpréter correctement. Le constructeur d'un test n'a pas seulement une responsabilité méthodologique, il doit également assumer une responsabilité éthique. L'instrument qu'il a créé va en effet servir à évaluer des sujets et à prendre des décisions à leur propos. Les informations communiquées dans le manuel doivent permettre de garantir un usage correct du test dans le respects des principes déontologiques.

2. La construction d'un test d'acquis scolaires

2.1 DÉFINITION DES FONCTIONS DU TEST

Dans l'enseignement, les tests sont appelés à jouer plusieurs rôles. L'instrument de mesure sera construit différemment selon la fonction à laquelle on le destine. Voici quelques usages courants des instruments de mesure en contexte scolaire :

1. dresser un bilan des acquis de l'élève ;
2. prendre une décision sur la promotion de l'élève ;
3. sélectionner les élèves selon certaines caractéristiques particulières afin de former des groupes ;
4. identifier les aspects de la résolution d'un problème source de difficultés ;
5. identifier les transferts qui ont ou n'ont pas eu lieu ;
6. préparer une révision de la matière à partir des points pour lesquels certains élèves éprouvent des difficultés ;
7. faire prendre conscience aux élèves de certains points majeurs de la matière.

Cette liste n'est pas exhaustive. Elle illustre simplement deux grands ensembles de situations où la mesure joue un rôle important en situation scolaire :

- *l'évaluation sommative* (situations 1, 2 et 3) ;
- *l'évaluation formative* (situations 4, 5, 6 et 7).

Dans le premier cas, on cherche à construire un instrument de mesure qui permette d'évaluer un échantillon de toute la matière enseignée. Un bon bilan nécessite un

échantillonnage du contenu qui soit exhaustif et représentatif. Pour ce faire, une mesure fondée sur les objectifs d'apprentissage est nécessaire.

Dans le second cas, on cherche à construire un outil qui permette une prise d'information focalisée et compréhensive. En fait, l'intérêt n'est pas de couvrir toute la matière, mais un aspect bien particulier de celle-ci. Alors que plusieurs objectifs peuvent être couverts dans un bilan, l'évaluation formative peut ne porter que sur un seul objectif. L'évaluation formative a pour fonction de fournir à l'enseignant et à l'élève une information pertinente sur le déroulement des apprentissages. C'est pourquoi l'enseignant veut contrôler plusieurs aspects de la tâche qu'il soumet à l'élève pour tester la stabilité des apprentissages dans différents contextes. C'est ce qu'une mesure critériée lui permet d'accomplir.

Tableau 1 – Fonctions de l'évaluation et qualités attendues des instruments de mesure

But de l'évaluation	Qualités souhaitées des mesures	Procédure
Obtenir des feed-back, observer	Informatives	Échanges spontanés, interrogations par essai-erreur
Faire un bilan, certifier	Représentatives et fiables	Définition des objectifs et tableau de spécification
Remédier aux difficultés, aider	Informatives, pertinentes et précises	Construction de tests critériés

Le tableau 1 décrit les différentes catégories de prise d'information que l'on rencontre en situation d'apprentissage scolaire. Les qualités de l'instrument de mesure doivent s'accorder avec le type d'information recherchée et les buts poursuivis. Dans bien des cas, une simple interrogation orale peut suffire. Dans des situations plus complexes où l'on doit articuler un plan d'intervention, cette prise d'information devra être complète pour rendre possible l'élaboration de stratégies d'enseignement adaptées (évaluation formative). Mais tout dépend de ce que l'on entend par information complète. Dans le cas d'un bilan (évaluation sommative), elle signifie que l'instrument de mesure couvre la totalité des contenus scolaires. Dans le cas d'une évaluation formative (ou diagnostique), elle signifie que l'instrument de mesure couvre l'ensemble des processus d'apprentissage pertinents. Dans les sections suivantes, la méthodologie utilisée pour construire des instruments sommatifs (§2.2) et diagnostiques (§2.3) sera détaillée.

2.2 L'ÉVALUATION SOMMATIVE

2.2.1 *La mesure fondée sur les objectifs*

Pour dresser un bilan représentatif des apprentissages des élèves, il faut que celui-ci reflète les objectifs du programme d'étude et de l'enseignement en salle de classe. Les programmes d'étude comportent généralement plusieurs catégories d'objectifs. Ceux-ci peuvent être regroupés selon leur spécificité (objectif global, général, spécifique) ou selon leur situation dans une séquence d'apprentissage (objec-

tif intermédiaire ou terminal). Quelle que soit la catégorie à laquelle il appartient, l'objectif possède des caractéristiques essentielles et des caractéristiques accessoires (tableau 2)

Tableau 2 – Formulation des objectifs d'apprentissage

	Obligatoire		Optionnel
Verbe	• un seul verbe • un verbe d'action • doit décrire un comportement univoque	**Contexte**	• ce qui est ou n'est pas disponible
Contenu	• un seul contenu par objectif • doit être un élément ou un sous-élément d'un programme	**Critères d'évaluation**	• condition d'acceptation de la performance • seuil de performance

Lors de la rédaction d'un objectif, les deux caractéristiques essentielles sont :

- un verbe d'action et un seul ;
- un contenu (complément d'objet) et un seul.

Le verbe d'action doit décrire un comportement observable directement (p. e. : cocher, souligner, écrire, lancer, etc.) ou indirectement (p. e. : identifier, choisir, etc.). Il ne doit y avoir qu'un seul verbe par objectif, sinon les attentes exprimées peuvent donner lieu à interprétation. Prenons l'exemple de la question suivante :

« Identifier et nommer les capitales provinciales du Canada »

La présence de deux verbes rend confuses les attentes en ce qui concerne les apprentissages des élèves. Sera-t-on satisfait lorsque l'élève saura nommer les capitales du Canada ou encore lorsqu'il pourra les identifier à partir d'une liste ou d'une carte géographique ? Pour considérer cet objectif comme atteint, faudra-t-il que l'élève manifeste les deux comportements (identifier et nommer) ou un seul des deux (identifier ou nommer) ? L'objectif manque de précision, non seulement à cause de l'ambiguïté créée par la présence de deux verbes, mais aussi parce que l'on ignore tout des conditions de réalisation de la performance et du seuil de réussite permettant de déterminer quand l'objectif peut être considéré comme atteint. Pour accroître la spécificité des objectifs, on ajoutera généralement les composantes suivantes à l'objectif :

- le contexte dans lequel sera réalisée la performance attendue ;
- le critère d'acceptation de la performance ;
- le seuil d'acceptation de la performance.

On ne s'attend pas à retrouver ces caractéristiques accessoires parmi les objectifs généraux. Par contre, elles sont essentielles à des objectifs dits spécifiques. La

figure 1 fournit un exemple d'un objectif spécifique comportant toutes ces composantes accessoires.

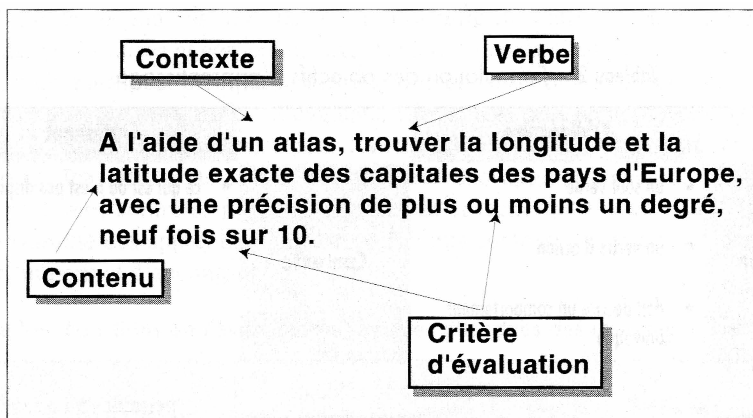


Figure 1 – Exemple de formulation d'un objectif

Le contexte décrit dans quelles conditions l'élève réalisera sa performance et ce qui sera à sa disposition. Dans le cas de l'exemple de la figure 1, il s'agit d'un atlas. Dans le cas d'autres objectifs, il pourrait s'agir d'une calculatrice (mathématiques), d'un dictionnaire ou d'une grammaire (français langue maternelle ou langue seconde). Le critère d'acceptation de la performance décrit le niveau de qualité de la performance attendue. Dans l'exemple, les coordonnées devront être relevées avec une précision d'un degré. Une erreur supérieure à un degré invaliderait la réponse en entier. Enfin, le seuil de réussite fournit un critère quantitatif pour considérer l'objectif comme atteint. Il établit combien de fois l'élève doit répéter sa performance au critère d'acceptation fixé pour que l'on considère qu'il maîtrise le contenu de l'objectif. Les seuils les plus courants oscillent généralement entre 80% et 100%. Dans le cas de l'objectif de la figure 1, ce seuil est de 90%. Qu'est-ce qui constituerait un seuil de réussite acceptable pour l'objectif « *identifier les capitales provinciales du Canada ?* », 80% ? 90% ? Cela pourrait dépendre des élèves à qui s'adresse cet objectif : le seuil pourrait être moindre pour des élèves belges que pour des élèves canadiens, par exemple.

Il existe plusieurs façons de déterminer un seuil de réussite. Cette question sera abordée plus en détail dans le chapitre 5. Pour l'instant, mentionnons que les composantes accessoires des objectifs sont parfois précisées dans les programmes d'étude en fonction des niveaux d'enseignement. Si elles ne sont pas précisées, elles peuvent souvent être déduites à partir d'informations complémentaires (par exemple, au Québec, les guides pédagogiques et les guides d'évaluation) et à partir du jugement de l'enseignant. Le contexte, le seuil et les conditions d'acceptation de la performance sont également des moyens de graduer les attentes en termes d'exigences et d'établir une progression dans les apprentissages. Ils permettent d'assurer une certaine continuité dans l'enseignement par objectif.

2.2.2 Le modèle de Deno et Jenkins et les taxonomies d'objectifs

Les objectifs spécifiques nous permettent de préciser la forme que prendra l'évaluation des apprentissages et les attentes que nous avons envers les élèves. Toutefois, ils sont peu pratiques pour considérer un programme d'étude dans son ensemble. Lorsqu'il s'agit de planification à long terme de l'enseignement et d'intégration des matières, les objectifs spécifiques peuvent devenir encombrants. L'intérêt doit alors se porter sur l'organisation des grandes parties de la matière et sur les processus cognitifs en jeu dans les apprentissages visés.

Deno et Jenkins (1969) ont élaboré un modèle qui tient compte de la spécificité nécessaire des objectifs à différents niveaux d'intervention. La figure 2 décrit les quatre niveaux (A à D) du modèle, allant de l'objectif global à la tâche d'examen. Il s'agit d'un modèle hiérarchique où chaque niveau supérieur contient les objectifs des niveaux inférieurs.

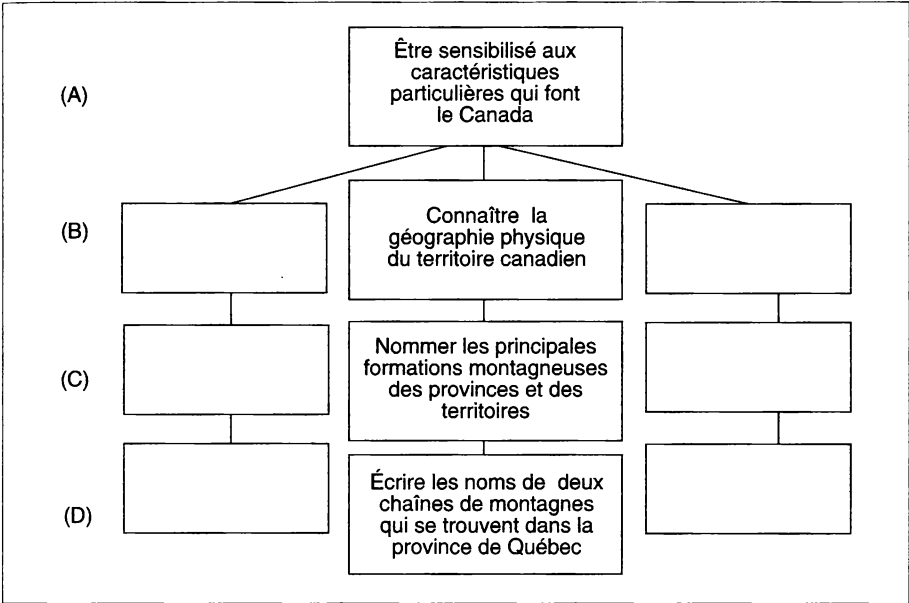


Figure 2 – Le modèle de Deno et Jenkins

Le niveau A est celui des *objectifs globaux*. Il sert à préciser les choix politiques, institutionnels, les grandes lignes du projet éducatif et de la mission de l'enseignement. Le niveau B cherche à préciser les objectifs globaux en situant le type de connaissances et d'habiletés (au niveau cognitif) ou le degré d'intériorisation (au niveau affectif) de l'objectif : c'est le niveau des *objectifs généraux*. Il ne s'agit pas à ce niveau d'indiquer de façon précise les attentes vis-à-vis des élèves. Il s'agit plutôt d'une première indication du degré d'approfondissement visé, tant au niveau cognitif qu'au niveau affectif. Les objectifs généraux sont particulièrement utiles pour dresser les grandes lignes d'un programme d'étude et articuler entre eux des objectifs qui peuvent, par leur nature et leur contenu, être fort différents. Au niveau C, les intentions se précisent à un tel point qu'on peut y indiquer les conditions précises d'évaluation :

catégorie de comportements attendus, contenus précis, conditions de réalisation de la performance et condition d'acceptation de la performance. C'est le niveau des *objectifs dits spécifiques*. Enfin, au niveau D, on retrouve les *tâches d'examen* elles-mêmes. C'est le niveau le plus spécifique des quatre niveaux du modèle. Ce n'est pas à proprement parler un objectif mais, comme le mentionne Ebel (1956), la tâche d'examen est la meilleure manière de connaître comment s'opérationnalisent les objectifs pédagogiques.

Le modèle de Deno et Jenkins permet de catégoriser les objectifs en fonction de leur spécificité, mais aussi en fonction de leur rôle dans un programme d'étude. Les objectifs spécifiques (niveau C) permettent de préciser ce qui sera évalué. Les objectifs généraux (niveau B) articulent les différents contenus d'un programme d'étude et précisent les processus visés par chaque grande catégorie d'apprentissage.

C'est au niveau B qu'interviennent les taxonomies d'objectifs généraux. On distingue trois grandes catégories taxonomiques :

1. objectifs cognitifs (Bloom, 1956) ;
2. objectifs affectifs (Krathwohl, 1964) ;
3. objectifs psychomoteurs (Harrow, 1972).

Dans le cas des objectifs cognitifs, l'objectif général permet de définir de manière suffisamment précise les connaissances et habiletés visées par le programme d'étude. La taxonomie des objectifs cognitifs de Bloom fait la distinction entre six niveaux d'habileté et d'acquisition de connaissances. Ces six niveaux hiérarchiques sont décrits dans la figure 3 (connaissances) et à la figure 4 (habiletés). La taxonomie des objectifs cognitifs joue ainsi un double rôle :

1. au niveau des programmes d'étude ;
2. au niveau de l'évaluation des apprentissages.

Au niveau des programmes d'étude, la taxonomie apporte plus de rigueur dans la définition de ce que l'on entend généralement par « connaissance », « compréhension » etc. De plus, elle permet de s'assurer que les attentes vis-à-vis les apprentissages des élèves sont conformes à leurs capacités et à leur développement cognitif. On peut ainsi établir une progression des habiletés intellectuelles impliquées dans l'apprentissage de mêmes contenus, mais à des niveaux scolaires différents. Par exemple, *« établir une classification du contenu de son herbier à partir d'un modèle fourni par l'enseignant »* constitue un objectif cognitif différent de celui qui consiste à *« élaborer une classification originale du contenu de son herbier à partir des échantillons de plantes recueillies »*. Le premier objectif porte sur l'application du modèle de l'enseignant, alors que le second repose davantage sur la synthèse (élaboration d'un plan d'action).

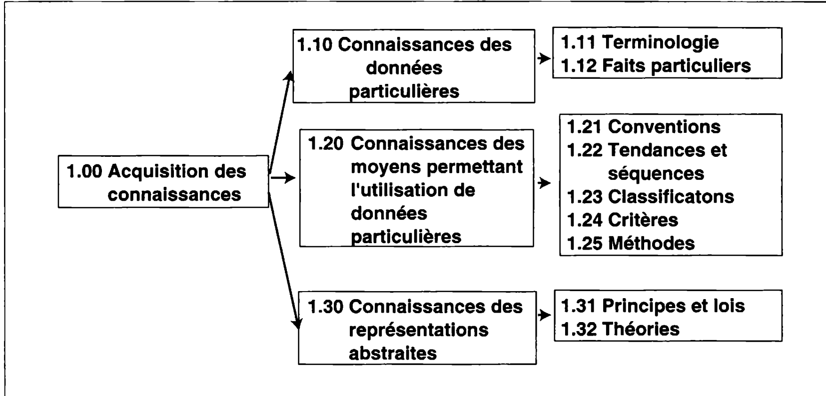


Figure 3 – Taxonomie des objectifs cognitifs : les connaissances

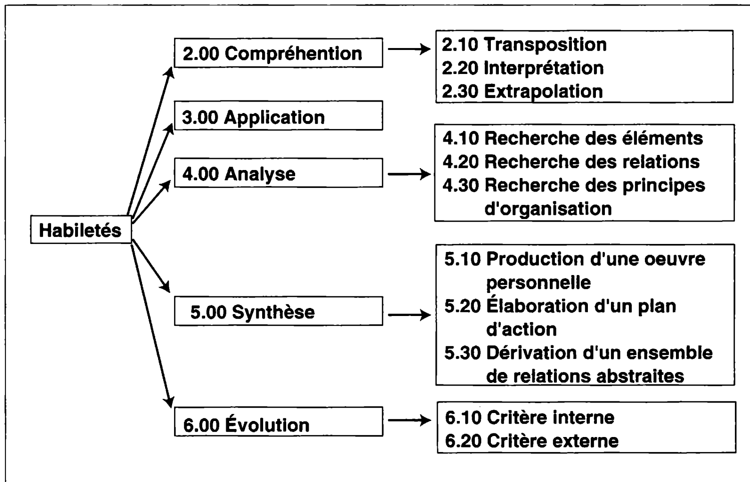


Figure 4 – Taxonomie des objectifs cognitifs : les habiletés

Au niveau de l'évaluation des apprentissages, la taxonomie permet de s'assurer que les processus de pensée activés lors de l'apprentissage seront mesurés lors de l'examen. Bloom et ses collaborateurs ont élaboré leur taxonomie après avoir constaté que les examens produits par les enseignants portaient habituellement sur la seule restitution des connaissances. Selon Bloom (1956), il était important de mesurer autre chose que les processus de pensée faisant intervenir principalement la mémorisation. Malheureusement, quelques trente années plus tard, Bloom (1984) constate que la situation n'a guère changé et que peu d'enseignants s'efforcent de mesurer les habiletés supérieures.

Les objectifs généraux ont également une incidence directe sur l'interprétation des objectifs spécifiques et, par ricochet, sur l'évaluation des apprentissages. Prenons une situation concrète assez répandue. Supposons que nous demandions à un étudiant de « *fournir un exemple de renforcement positif* ». S'il s'agit d'un objectif de connaissance, il suffira à l'étudiant de répéter un exemple qu'il a entendu en classe ou lu dans

le manuel obligatoire du cours. Si, par contre, il s'agit d'un objectif de compréhension, nous nous attendons à ce que l'étudiant fournisse un exemple original. La réitération d'un exemple connu ne serait pas suffisante pour parler de compréhension. De ce dernier exemple, nous pouvons conclure qu'une même tâche peut être employée pour mesurer des niveaux taxonomiques fort différents. La condition d'acceptation de la performance permet dans ce cas-ci de s'assurer que la question d'examen mesure bien le niveau taxonomique qu'il est censé mesuré. Pour que les choses soient claires pour l'élève, il faudra que l'énoncé de la question soit sans équivoque à propos de cette condition d'acceptation. Par exemple :

> *« Écrivez un exemple original de renforcement positif. Les exemples du manuel de cours ou du professeur ne seront pas acceptés ».*

2.2.3 *Objectifs terminaux et objectifs intermédiaires*

Dans un autre ordre idée, il est parfois nécessaire d'aborder l'articulation des objectifs dans la séquence d'apprentissage. La taxonomie des objectifs permet de décrire une hiérarchisation des processus de pensée, mais cette articulation est fort générale. De plus, le type de relation décrite par la taxonomie des objectifs se limite à l'inclusion. D'autres relations entre objectifs d'apprentissage sont possibles.

Lorsque l'on souhaite préciser l'enchaînement de plusieurs objectifs dans un programme d'études, on peut distinguer les objectifs terminaux des objectifs intermédiaires. L'objectif terminal décrit la finalité ultime d'un apprentissage, son point d'arrivée. L'objectif intermédiaire énumère les étapes nécessaires qui doivent jalonner le cheminement de l'élève du point de départ au point d'arrivée. Sans la maîtrise de ces jalons, la maîtrise de l'objectif terminal est compromise. Par contre, lorsque l'objectif terminal est atteint, on peut conclure que les objectifs intermédiaires ont été maîtrisés.

Les objectifs terminaux conviennent particulièrement à l'évaluation sommative. Ils permettent de couvrir une grande étendue de contenu. De plus, il est normal qu'un bilan porte sur les apprentissages complétés plutôt que sur ceux qui sont en voie de réalisation. Enfin, lorsqu'il s'agit d'établir un bilan, il est généralement trop tard pour se demander à quel moment de l'apprentissage l'étudiant a éprouvé des difficultés. Par contre, cette dernière information peut être utile dans le cas d'une évaluation formative ou encore de ce qu'il est convenu d'appeler une évaluation « micro-sommative » (Scallon, 1992). Afin de mieux comprendre les raisons d'une difficulté au niveau d'un objectif terminal, il peut alors être utile de s'assurer que tous les prérequis sont bien maîtrisés. Le degré de maîtrise de chaque objectif intermédiaire peut nous renseigner sur les moyens de corriger une difficulté.

2.2.4 *Échantillonnage des items et tableau de spécification*

Certains instruments de mesure, en particulier les examens, doivent être administrés à période fixe afin de dresser un bilan des apprentissages de l'élève. Cette évaluation ne répond à aucun besoin particulier de la part de l'enseignant ou de l'élève, mais elle correspond à une exigence administrative. Ceci ne signifie pas que l'enseignant ne soit pas intéressé de temps à autres à effectuer un bilan des apprentissages de ses élèves pour son propre compte. Mais ce bilan se ferait probablement de façon fort différente. Par exemple, l'enseignant pourrait décider d'éliminer de tels bilans les

items qu'il considère comme réussis depuis longtemps par une grande majorité des élèves. Pour certifier un cycle d'apprentissage, cependant, la couverture de la matière devra être exhaustive, même si elle porte sur des points pour lesquels l'enseignant est déjà assez bien informé.

Le bilan, qu'il corresponde à une exigence administrative, se doit d'être représentatif. Ce qui est représentatif peut différer selon l'usage qui sera fait du bilan en question. Lorsqu'il s'agit de certification, cette définition doit être stricte. L'enseignant a peu de marge de manoeuvre quant à l'univers des situations qu'il peut échantillonner pour son examen. Afin d'assurer la comparabilité des résultats entre classes, les enseignants de cinquième primaire, par exemple, devront tirer leurs questions d'examen d'un même ensemble. Ce ne seront pas les mêmes questions, mais elles devraient, dans la mesure du possible, constituer des ensembles parallèles facilement comparables et congruents avec le programme d'étude commun à tous les élèves.

L'échantillonnage est l'un des outils à la disposition de l'enseignant pour construire son instrument de mesure. Tout comme l'échantillonnage des sujets (voir chapitre 7), l'échantillonnage des questions peut prendre plusieurs formes :

1. *Échantillonnage aléatoire simple.* Chaque question a une chance égale d'être choisie.
2. *Échantillonnage stratifié.* Le test entier comporte des questions appartenant à un objectif dans une proportion qui correspond à l'importance de cet objectif dans la matière à couvrir.
3. *Échantillonnage par grappes.* L'échantillonnage, dans ce cas, ne se fait pas par question, mais par objectif, car le nombre d'objectifs à couvrir est extrêmement grand.
4. *Échantillonnage double.* L'échantillonnage se fait en deux étapes : (a) d'abord les objectif et ensuite (b) les questions à l'intérieur des objectifs.

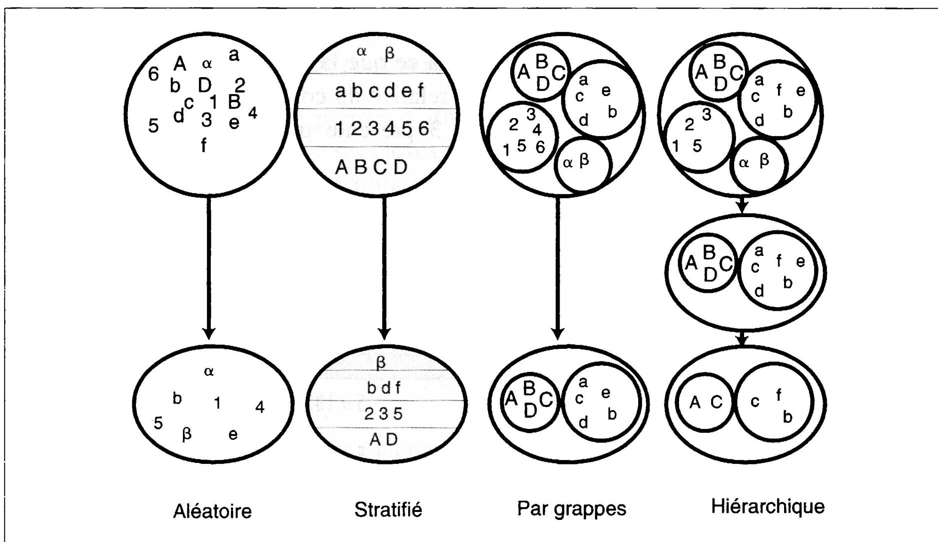


Figure 5 – Techniques d'échantillonnage des questions

Ces méthodes d'échantillonnage sont décrites au moyen des quatre schémas à la figure 5. La méthode aléatoire simple signifie que chaque item a une chance égale d'être choisi. La méthode stratifiée est également une méthode aléatoire. Elle est particulièrement utilisée lorsque le nombre d'items à choisir au départ est relativement faible et que l'on veut s'assurer que les items se retrouveront dans notre échantillon dans les mêmes proportions que dans le domaine d'où ils ont été tirés. Par exemple, si 33% des exercices faits en classe ont porté sur la physiologie et 50% sur l'anatomie, l'examen devrait refléter cette distribution. La méthode par grappes et le méthode hiérarchique impliquent une sélection des objectifs. Lorsqu'un objectif n'est pas choisi au hasard, aucun item concernant cet objectif ne se retrouve dans l'examen. Dans la méthode par grappes, tous les items touchés par les objectifs choisis seront retenus. Dans la méthode hiérarchique, un choix au hasard des items parmi les objectifs déjà choisis permettra d'en restreindre le nombre total. Cette dernière méthode d'échantillonnage des items s'avère particulièrement utile lorsque le contenu de la matière à couvrir est fort vaste.

Il est important de noter que seules les deux premières techniques d'échantillonnage permettent, avec un nombre suffisamment grand d'items, d'échantillonner toute la matière. Avec les deux derniers types d'échantillonnage, certaines parties de la matière seront nécessairement omises. Cet inconvénient n'est pas majeur lorsqu'il s'agit d'un examen qui fait suite à une série d'examens partiels. Cette méthode d'échantillonnage est caractéristique des examens de fin d'année. Par contre, les bilans plus fréquents (fin d'étape) ne peuvent omettre complètement un objectif.

Le tableau de spécification est un moyen utilisé depuis longtemps pour s'assurer que l'échantillonnage des questions d'examen est véritablement représentatif de la situation qui a prévalu en salle de classe ou encore des exigences décrites dans le programme d'étude. Il prend généralement la forme d'un tableau de contingence à double entrée, la première étant constituée du contenu, la seconde du niveau taxonomique des objectifs mesurés. Un grand soin est accordé à ce que la proportion des items d'examen corresponde étroitement à l'importance relative du contenu et du niveau taxonomique du programme d'étude. Le tableau 3 présente un exemple de tableau de spécification pour un examen de géographie.

Tableau 3 – Exemple de tableau de spécification

	Niveau taxonomique		
	Connaissance	**Compréhension**	**Total**
Géographie humaine	10 = 20%	5 = 10%	15 = 30%
Géographie politique	10 = 20%	5 = 10%	15 = 30%
Géographie physique	10 = 20%	10 = 20%	20 = 40%
Total	30 = 60%	20 = 40%	50 = 100%

Le tableau de spécification correspond à un échantillonnage stratifié. Dans l'exemple de l'examen de géographie du tableau 3, la stratification s'est effectuée en tenant compte du contenu (géographie humaine, politique ou physique) ainsi que du niveau taxonomique (connaissance, compréhension). En principe, la répartition des items d'examen selon ces deux caractéristiques doit refléter l'importance consacrée en classe, en terme de temps d'étude ou d'enseignement. Si 10% du temps en classe a été consacré à la compréhension de la géographie politique, 10% des 50 questions d'examen (5 questions) devraient porter sur cette matière. À défaut de trouver autant de questions, il est toujours possible d'ajuster la pondération de l'examen de manière à rendre plus représentatif le score total. L'alternative à cinq questions d'un point chacune, pourrait ainsi être une question de deux points et une autre de trois points sur la géographie politique.

D'autres caractéristiques que le niveau taxonomique ou le contenu peuvent être employées pour établir un tableau de spécification. Le type de production (convergente ou divergente), le format d'items (choix de réponse ou réponse élaborée) peuvent également entrer en considération. Néanmoins, l'exemple précédent est sans doute plus représentatif de ce qui se passe en contexte scolaire. En effet, l'organisation habituelle des programmes d'étude favorise plutôt ce genre de stratification.

2.3 L'ÉVALUATION CRITÉRIÉE

2.3.1 *Introduction*

La mesure critériée regroupe un ensemble de procédures permettant une prise d'information détaillée à propos de l'apprentissage d'un sujet. Ces procédures ont en commun de mieux définir et de mieux contrôler les critères quantitatifs et qualitatifs de la performance, tels que :

- les aspects de la présentation d'une tâche ;
- les conditions de réalisation d'une tâche ;
- les niveaux d'exigence pour la réalisation d'une tâche.

La mesure critériée permet d'affiner la prise d'information de l'enseignant à propos des apprentissages de ses élèves et le rend ainsi plus apte à comprendre les raisons de leurs difficultés. La planification de l'enseignement en est dès lors facilitée. Plusieurs techniques de spécification de domaine permettent de construire des instruments de mesure critériée. Voici une liste de techniques que nous allons présenter de manière détaillée :

- l'objectif enrichi ;
- l'analyse des concepts ;
- la théorie des facettes.

Il existe plusieurs autres techniques de mesure critériée (Roid et Haladyna, 1982). Chacune se réfère à une conception particulière de ce qu'est instrument de mesure. Il est donc nécessaire de se familiariser avec plusieurs de ces techniques si l'on veut être capable d'employer adéquatement la mesure critériée dans une grande variété de situations didactiques.

L'objectif spécifique donne souvent lieu à une telle marge d'interprétation dans la rédaction des tâches d'examen qu'il devient difficile de considérer celles-ci comme appartenant au même domaine. Prenons l'exemple de l'objectif spécifique suivant : *« À l'aide de la règle, mesurer les dimensions d'une figure géométrique »*. Plusieurs situations fort différentes peuvent être construites pour vérifier la maîtrise de cet objectif. Considérons les facteurs qui peuvent intervenir :

- le type de figure géométrique : parallélogramme, triangle, cercle ;
- la nature de la dimension : explicite (le côté d'un carré, d'un triangle) ou implicite (la diagonale d'un carré ou la hauteur d'un triangle dans certains cas) ;
- l'orientation de la figure dans l'espace plan ;
- les caractéristiques particulières de la figure : le type de quadrilatère (carré, rectangle, losange, parallélogramme, trapèze) ; le type de triangle (équilatéral, isocèle, rectangle, scalène, etc.).
- la quantité et le type d'information fournis au départ.

Dans le cas du triangle, on peut imaginer une diversité de situations mettant en oeuvre cette tâche. La figure 6 présente une série d'items basés sur le même objectif. Toutes ces items sont parfaitement congruents avec l'objectif de départ, mais, de manière évidente, chaque item fait intervenir des habiletés fort différentes, selon le type de triangle choisi.

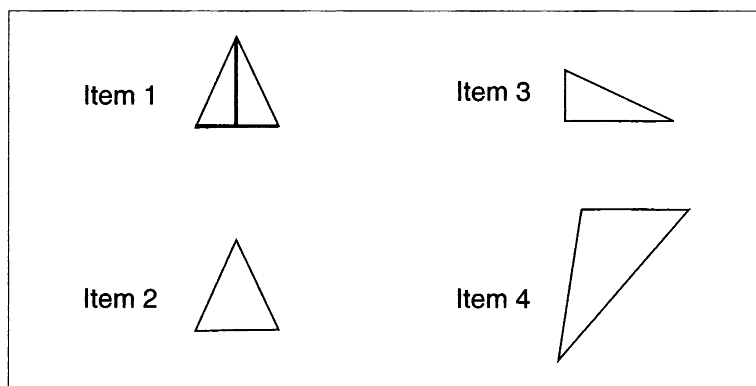


Figure 6 – Exemple d'items se référant au même objectif

L'item 1 porte sur un triangle pour lequel une hauteur et une base sont déjà tracées sans indiquer quel segment de droite est en fait la base et quel segment est la hauteur. Un tel exercice permet de déterminer si l'élève différencie la base de la hauteur et s'il sait prendre ses mesures de manière adéquate. L'item 2 laisse le soin à l'élève d'identifier lui-même base et hauteur. Toutefois, ce triangle ne présente pas de difficulté particulière comme les deux qui suivront. Il serait difficile de généraliser que l'élève sait mesurer la base et la hauteur d'un triangle à partir d'items comme le n° 3. Celui-ci présente un cas particulier de triangle : le triangle rectangle. Dans ce triangle, deux bases et deux hauteurs (correspondants aux deux angles non droits) correspondent à l'un des côtés de l'angle droit. Ce type d'exercice présente une difficulté particulière qui permet de mesurer le degré de généralisation des notions de base et de

hauteur. L'exercice 4 présente un triangle scalène dans lequel une des bases se situera à l'extérieur du triangle. Il est important de soumettre à l'élève des exemples de ce type pour s'assurer que l'objectif d'apprentissage est atteint dans toutes les situations, notamment celles où la base ne se situe pas à l'intérieur du triangle. Si les élèves ont été habitués uniquement à prendre des mesures sur des figures telles que celles des items 1 et 2, les items 3 et 4 risquent de les dérouter. Par contre, s'ils ont été amenés à véritablement comprendre les concepts de base et de hauteur, ce changement de caractéristiques du contenu ne devrait pas être source de difficultés particulières et ils devraient facilement généraliser leurs apprentissages.

Il existe donc différentes façons de concevoir des tâches mesurant l'atteinte de l'objectif *« À l'aide de la règle, mesurer la base et la hauteur d'un triangle »*. Certaines mettent l'accent sur l'action de mesurer (la base et la hauteur étant identifiées au départ), d'autres sur la compréhension des concepts (trouver la base et la hauteur à mesurer). L'interprétation des résultats est donc susceptible de changer selon le type de situation à laquelle on expose l'élève et selon les conditions dans lesquelles s'est effectué l'apprentissage.

2.3.2 *L'objectif enrichi*

C'est sans doute la technique la plus facile à apprendre, une fois que l'on connaît bien la mesure fondée sur les objectifs. Élaborée par Popham (*« amplifed objectives »*), cette technique de spécification de domaine a pour but de pallier les limites de l'objectif spécifique en en fournissant une description enrichie. L'objectif enrichi permet de réduire les possibilités d'interprétation en définissant l'objectif avec plus de rigueur. Popham (1980) a défini l'objectif enrichi en distinguant trois parties principales :

1. l'énoncé de l'objectif.
2. un exemple d'item incluant :
   - une directive ;
   - un exemple.
3. un complément d'information sur :
   - l'examen (ce que l'élève aura à faire lors du test, la nature du stimulus) ;
   - les choix de réponse ;
   - les critères de correction.

La figure 7 présente une exemple d'objectif enrichi qui permet de mieux spécifier le domaine des apprentissages et de mettre de l'ordre dans les différentes situations décrites par les items 1 à 4 de la figure 6.

Objectif : A l'aide d'une règle, mesurer la longueur de la base et de la hauteur d'un triangle

Exemple d'item :

Voici une série de triangles. Sous chaque triangle, indiquez la longueur de la base et la hauteur en millimètres. La base est le côté du triangle tracé en gras.

	(A)	(B)	(C)	(D)
Base	.... mm	.... mm	.... mm	.... mm
Hauteur	.... mm	.... mm	.... mm	.... mm

Conditions de réalisation de la performance :

1. les triangles sont quelconques ;
2. les triangles sont diversement orientés ;
3. la base est tracée en gras ;
4. le côté désignant la base est déterminé au hasard ;
5. la hauteur n'est pas identifiée ;
6. le sujet dispose d'une règle graduée en centimètres et en millimètres.

Critères de cotation :

1. le sujet inscrit sa réponse dans la case appropriée ;
2. la réponse du sujet doit être exacte à 1 millimètre près.

Figure 7 – Exemple d'un objectif enrichi

À partir de l'objectif enrichi décrit dans la figure 7, le praticien peut rédiger un grand nombre de questions. Chacune de ces questions appartiendra au même domaine et il sera possible d'obtenir une mesure absolue et précise de la maîtrise de l'objectif. Supposons en effet qu'un enseignant prépare 10 items à partir de la définition précédente de l'objectif enrichi. Il n'y a pas de raison de supposer que le test ainsi construit sera plus facile ou plus difficile que celui construit par un autre enseignant à partir de la même description. De plus, si un élève réussit 80% des items de ce domaine, il n'y aucune raison de supposer qu'il ne pourra atteindre le même score avec un autre échantillon d'items tirés du même domaine, tel que spécifié par l'objectif enrichi.

L'objectif enrichi nous permet donc de nous prononcer avec une plus grande assurance sur le degré de maîtrise et de non maîtrise d'un objectif. En effet, tout échantillon d'items servant à mesurer la maîtrise de l'élève provient du même domaine et les chances de variation d'échantillonnage d'un enseignant à un autre sont réduites au minimum.

2.3.3 *L'analyse de concepts*

Lorsqu'il s'agit de mesurer la maîtrise d'un concept, le praticien peut souhaiter déterminer le degré de discrimination que le sujet réussit à atteindre entre le concept étudié et les concepts voisins. Il peut aussi chercher à déterminer dans quelle mesure l'apprentissage d'un nouveau concept contribue à changer la représentation initiale du sujet ou encore une représentation erronée (ou pré-concept). Le praticien peut également vouloir déterminer à quel point le sujet est capable de généraliser un concept appris à l'ensemble des situations auxquelles il peut s'appliquer.

Dans le cas précis de la hauteur d'un triangle, plusieurs facteurs peuvent contribuer à ce qu'un élève ait une mauvaise représentation du concept. C'est pourquoi il est important qu'il soit capable de faire la différence entre les caractéristiques essentielles et les caractéristiques accessoires du concept étudié. L'analyse de concepts contribue à spécifier un domaine d'items servant à tester l'apprentissage de l'élève. Le tableau 4 présente un exemple d'analyse du concept « hauteur d'un triangle ».

Tableau 4 – Exemple d'analyse du concept « hauteur d'un triangle »

Caractéristiques essentielles

1. Segment de droite
2. Relie un sommet du triangle au côté opposé (base)
3. Fait un angle droit avec le côté opposé à l'un des sommets du triangle

Caractéristiques accessoires

1. Le segment de droite peut être (1) intérieur, au triangle, (2) extérieur au triangle, (3) un de ses côtés
2. L'orientation d'un triangle n'a aucun effet sur sa hauteur ; la base peut être (1) horizontale, (2) verticale, ou (3) oblique
3. Le type de triangle : (1) équilatéral, (2) isocèle, (3) rectangle, ou (4) scalène

L'analyse des concepts comporte cinq parties :

1. la définition des caractéristiques essentielles ;
2. la définition des caractéristiques accessoires ;
3. une série d'exemples et de contre-exemples tirés de l'enseignement ;
4. une série d'exemples et de contre-exemples pour l'évaluation (similaires à ceux de l'enseignement) ;

L'analyse des concepts permet de s'assurer que les items porteront sur des situations similaires à celles vues au cours : la congruence entre l'évaluation et l'enseignement est ainsi assurée. Elle permet aussi, si l'enseignant le désire, de spécifier un ensemble de situations, légèrement différentes de celles vues en classe, afin de vérifier

s'il y a généralisation des apprentissages. Mais, il doit s'agir là d'un objectif bien particulier. Il n'est pas équitable de mesurer ce genre d'habileté sauf si l'enseignant a présenté en classe certaines des généralisations possibles à l'aide d'autres exemples et contre-exemples.

L'analyse des concepts fait plus que préciser le domaine des items. Elle permet aussi d'envisager certaines erreurs conceptuelles qui peuvent être fort utiles lorsqu'il s'agit de rédiger des leurres pour des questions à choix multiple. Ainsi, l'analyse des leurres permet d'identifier de manière plus précise le type de difficulté de l'élève. Cette caractéristique particulière de l'analyse des concepts lui confère un avantage certain sur l'objectif enrichi pour certains types d'évaluation.

2.3.4 *La théorie des facettes*

Guttman (1969) a élaboré la *théorie des facettes* afin de nous permettre d'exercer un meilleur contrôle sur les caractéristiques des items. La théorie des facettes a d'abord été employée pour la mesure des attitudes, mais depuis, son usage a été généralisé à la mesure des apprentissages.

Tableau 5 – Domaine d'items d'addition défini selon trois facettes

Opération d'addition		Présentation verticale	Présentation horizontale
nombres à deux chiffres	sans retenue	11 + 34	81 + 12 = ...
	avec retenue	47 + 29	27 + 75 = ...
nombres à trois chiffres	sans retenue	252 + 127	523 + 110 = ...
	avec retenue	173 + 451	815 + 105 = ...
nombres à quatre chiffres	sans retenue	1342 + 2113	1177 + 2122 = ...
	avec retenue	1578 +8112	8722 + 1281 = ...

La spécification d'un domaine d'items à l'aide de la théorie des facettes est analogue au choix d'un plan d'observation lors d'une recherche expérimentale. Le praticien choisit les facettes d'intérêt en fonction de ses objectifs et détermine pour chaque facette un certain nombre d'éléments, ou valeurs, que la facette peut prendre. Le croisement de plusieurs facettes donne lieu à un grand nombre de possibilités d'items dont les résultats permettront de déterminer la maîtrise ou la non maîtrise d'une habileté selon les situations. Plusieurs profils de performance pourront ainsi être mis au point.

Le tableau 5 présente un exemple de test d'arithmétique (opération d'addition) construit selon des facettes

L'exemple présenté dans le tableau 5 illustre le domaine d'items d'addition défini selon trois facettes :

1. la présentation horizontale ou verticale
2. l'ordre de grandeur des nombres (2, 3 ou 4 chiffres) ;
3. le type d'addition (avec ou sans retenue).

Comme il y a deux éléments dans la facette 1, trois dans la facette 2 et deux dans la facettes 3, il y a (2 x 3 x 2) = 12 combinaisons possibles. La tableau 5 ne présente qu'un exemple d'items pour chaque interaction des différents éléments des trois facettes. On peut s'imaginer cependant la facilité qu'il y a à construire des items critériés équivalents sur base des facettes de ce tableau à double entrée.

Ce modèle de spécification de domaine est très pratique lorsque l'on souhaite établir un profil des performances d'un sujet dans différentes situation. Il est alors possible d'identifier le ou les éléments de la ou des facettes qui posent des difficultés au sujet. Le tableau 6 présente un exemple de profil que l'on peut déduire de la spécification de domaine du tableau 5.

Tableau 6 – Profil de performance basé sur l'analyse des facettes

	2 chiffres	3 chiffres	4 chiffres	total
avec retenue	....../2	....../2	....../2	....../6
sans retenue	....../2	....../2	....../2	....../6
total	....../4	....../4	....../4	....../12

En regroupant les résultats par facettes ou combinaisons de facettes, il est possible de mettre en évidence plusieurs profils de performance. L'un en fonction de l'ordre de grandeur de l'addition regroupe quatre items. L'autre selon le type d'algorithme (avec ou sans retenue) regroupe 6 items. Si un élève réussissait 6 items sur 6 « sans retenue » et 2 items sur 6 « avec retenue », on pourrait conclure à une difficulté au niveau de l'application de l'algorithme de retenue. De plus, on pourrait affirmer que la compétence à additionner des nombres « sans retenue » s'est généralisée à tous les nombres entiers, quelque soit leur ordre de grandeur.

Le praticien pourrait pousser plus loin l'analyse du résultat des additions avec retenue. Les deux additions réussies ont-elles une caractéristique en commun ? S'il

s'avère que, dans les deux cas, il s'agit de nombres à deux chiffres, alors le problème se pose non au niveau de l'algorithme d'addition avec retenue mais au niveau de sa généralisation à des situations où plus d'une retenue est possible. En effet, avec des nombres à trois et quatre chiffres, il est possible qu'il y ait deux retenues et même trois retenues. Cette facette du problème ne fait pas partie de la spécification du domaine d'items. Si elle s'avérait pertinente, elle pourrait être incluse dans un nouvel instrument d'évaluation.

3. Les formats d'items

3.1 FORMATS D'ITEMS POUR LES TESTS COGNITIFS

3.1.1 *Typologie des formats d'items*

La manière de classer les formats d'items varie d'un auteur à l'autre. Le tableau 7 présente une typologie classique distinguant essentiellement les questions ouvertes et les questions fermées. Les premières demandent aux sujets de produire la réponse. Les secondes demandent, elles, de faire un choix parmi un certain nombre d'alternatives déjà données. Comme nous le verrons dans le § 3.1.2, les formats fermés ont été développés pour répondre à un certain nombre de problèmes liés aux exigences de la théorie classique des tests. La nécessité de réduire la part de l'erreur dans la variance du score total a en effet conduit à standardiser au maximum les modalités de passation et de cotation des items. Aux yeux du grand public, les questions fermées sont d'ailleurs étroitement liées à la notion de test. Elles véhiculent avec elles un certain nombre de représentations, souvent fausses, qui appellent une mise au point.

Tableau 7 – Typologie des formats d'items

Questions fermées	Questions ouvertes
• Questions à choix multiple	• Questions à réponse brève
• Questions « vrai-faux »	• Questions à réponse narrative
• Questions d'appariement	• Questions demandant une performance

Au sein des questions fermées, certaines distinctions plus fines peuvent être faites en fonction du type de choix demandé aux sujets. Les *questions à choix multiple*, comme leur nom l'indique, demandent de réaliser un choix parmi plusieurs options. Ces questions comprennent deux parties : *une amorce* (ou prémisse), qui présente le problème, suivie de plusieurs *alternatives* (choix, options...) qui sont autant de solutions possibles au problème posé. Outre la solution correcte, les alternatives comprennent des solutions incorrectes, appelées *distracteurs*. Le nombre d'alternatives peut varier mais il est souvent limité à quatre choix.

EXEMPLE

La variance d'une distribution est une mesure de :

A. dispersion.

B. tendance centrale.

C. relation.

D. localisation.

Les *questions « vrai-faux »* sont, elles, plus simples dans leur présentation puisqu'elles comportent une seule proposition dont le sujet doit évaluer la véracité. Pour répondre, celui-ci doit entourer son choix « VRAI » ou « FAUX » (ou parfois, « OUI-NON », ou encore, « D'ACCORD-PAS D'ACCORD »).

EXEMPLE

Un item possède un indice de discrimination D de 0,8. Cela signifie que cet item est très discriminatif.

Vrai — Faux

Enfin, les questions d'appariement sont une forme dérivée des questions à choix multiple. Au lieu de construire quatre questions à choix multiple (ou plus encore), il peut être plus économique de ne retenir que les amorces de ces questions et les solutions correctes à celles-ci et de demander aux sujets de mettre correctement en relation les quatre amorces, appelées prémisses (colonne A de l'exemple), et les quatre réponses (colonne B).

EXEMPLE

La colonne A contient une liste de poèmes de la période romantique. La colonne B contient, elle, une série d'auteurs français de cette période. Associez chacun de ces poèmes avec son auteur. Pour ce faire, notez en face du poème, la lettre qui correspond au nom de son auteur.

Colonne A	*Colonne B*
1. Le lac	A. Victor Hugo
.... 2. A Villequier	B. Alfred de Musset
.... 3. Ballade à la lune	C. Alfred de Vigny
.... 4. La mort du loup	D. Alphonse de Lamartine

Bien que rangées parmi les questions ouvertes, les *questions à réponse brève* possèdent souvent des caractéristiques proches de celles des questions fermées. Certains auteurs les qualifient d'ailleurs de questions « objectives » (p.e. Ebel & Frisbie, 1991, p.179). En effet, elles demandent aux sujets de fournir un mot, une phrase ou un nombre qui peut être évalué comme correct ou incorrect, sans qu'intervienne la subjectivité du correcteur. Les exemples suivants illustrent bien le caractère objectif de telles questions.

EXEMPLE

A. Qui a découvert le vaccin contre la rage ?..

B. Combien de jours y a-t-il dans une année ? ....

C. Quelle est l'aire, en cm^2, d'un triangle dont la hauteur est de 16 cm et la base de 8 cm ? ....

Mais toutes les questions à réponse courte ne sont pas aussi objectives. Certaines demandent un jugement du correcteur à propos de la qualité de la réponse. Par exemple, si la question est : « Donnez un synonyme du mot *aimable* ». La réponse se réduira à un seul mot, mais sa qualité devra être appréciée par le correcteur. Dans un cas comme celui-ci, une certaine variabilité entre correcteurs peut être observée, ce qui diminue la fiabilité d'une telle question. Les problèmes liés à la subjectivité des correcteurs sont toutefois nettement plus importants avec les *questions à réponse narrative*. De telles questions offrent une grande liberté de réponse aux sujets. Ils peuvent en effet décider de la manière d'aborder le problème posé, du type d'informations à utiliser, de la façon d'organiser leur réponse et de l'accent à mettre sur les différentes parties de celle-ci. Les questions à réponses narratives sont, par conséquent, bien adaptées pour évaluer la capacité d'un sujet à organiser, à intégrer et à exprimer ses idées. Malheureusement, la richesse de l'information ainsi recueillie se paie par une complexité et une subjectivité importante de la cotation.

EXEMPLE

A. Comparez les conceptions de l'apprentissage de Piaget et de Skinner.

B. Comment Spitz explique-t-il les angoisses dites « du huitième mois » ?

C. Décrivez les étapes essentielles du développement au cours du stade sensori-moteur.

Les questions, ouvertes ou fermées, qui ont été présentées jusqu'ici se caractérisent par le rôle essentiel qu'y joue le langage, que ce soit au niveau des stimuli présentés au sujet, des processus mentaux que celui-ci met en oeuvre et des réponses qu'il produit. Les *questions de performance* font, elles, intervenir le langage à un degré nettement moindre. Elles demandent en effet aux sujets de réaliser une action où le langage peut être totalement absent (jouer un morceau de musique, construire un puzzle, dessiner des formes géométriques...). Par nature, ces questions sont « ouvertes » puisque le sujet doit produire la réponse. Elles font souvent intervenir un matériel plus ou moins standardisé afin de permettre des comparaisons et d'avoir un certain contrôle sur les critères d'évaluation des productions des sujets. Les questions de performance sont particulièrement intéressantes pour évaluer certaines compétences cognitives et certaines habiletés professionnelles.

Plusieurs habiletés typiques d'une profession ne peuvent être correctement évaluées qu'au travers des performances des sujets. Comment, par exemple, évaluer un musicien autrement que par une performance musicale ? Les performances demandées peuvent être identiques à celles produites dans le cadre professionnel, comme dans le cas de la performance musicale. Elles peuvent aussi reposer sur une situation professionnelle simulée. Par exemple, on peut demander à un candidat cadre d'entreprise de planifier une journée fictive à partir d'un ensemble de contraintes données par l'examinateur. Lorsque l'objectif est d'évaluer des habiletés professionnelles, l'accent est mis sur les productions du sujet, et non sur les processus mis en oeuvre pour réaliser ces productions. Le psychologue d'entreprise désire en effet vérifier si, par exemple, une candidate secrétaire peut dactylographier correctement une lettre et effectuer sans erreur un classement de documents. Il ne s'intéresse pas aux processus mentaux qu'utilise cette secrétaire pour parvenir au résultat. Par contre, dans les domaines cliniques et éducatifs, les questions de performance servent avant tout de révélateur de cer-

taines caractéristiques cognitives. Par exemple, la reproduction de dessins à l'aide de cubes colorés (Cubes de Kohs) peut servir à évaluer les capacités de raisonnement spatial. De même, la reproduction différée de dessins géométriques procure des informations utiles à propos du fonctionnement de la mémoire. Ici, la production du sujet n'a pas de valeur intrinsèque, elle ne sert que d'indicateur de capacités cognitives inaccessibles à l'observation directe. De tels items de performance sont surtout utilisés lorsque l'accès au langage est limité (jeunes enfants, sujets maîtrisant mal la langue de l'examinateur, adultes atteints de lésions cérébrales...) ou lorsque la composante verbale de la compétence cognitive visée est réduite (p.e. l'organisation de l'espace, la coordination oculomanuelle...).

3.1.2 *Question fermée ou question ouverte ?*

Le choix entre des questions ouvertes ou des questions fermées est souvent déterminé par les a priori plus que par une réelle connaissance de leurs propriétés respectives. De nombreux praticiens rejettent viscéralement les questions fermées, accusées de réduire l'apprentissage à une simple accumulation de connaissances, de négliger les compétences cognitives les plus élevées, d'encourager le « bachotage »... Certaines de ces critiques sont certes fondées, mais la plupart ne témoignent que du manque d'information de leurs auteurs.

En fait, il n'y a pas lieu de décider dans l'absolu de choisir des questions ouvertes ou des questions fermées. Aucun format n'est le meilleur « en général ». Le problème doit être posé en d'autres termes. La véritable question est en effet : *« quand faut-il utiliser tel ou tel format d'item ? »*. C'est en fonction des objectifs du test et de ses conditions d'application qu'un format peut être considéré comme le plus adéquat. Dans certains cas, des questions fermées seront plus appropriées alors que dans d'autres cas, des questions ouvertes seront plus adéquates. Avant de choisir un format d'item, le praticien doit envisager les différentes contraintes qui doivent être prises en compte. Le choix final correspondra au meilleur équilibre entre ces différentes contraintes. Celles-ci peuvent être rangées en quatre catégories que nous allons détailler.

(1) Les capacités cognitives à mesurer

Les questions fermées ont la réputation de ne permettre d'évaluer que les niveaux les plus bas de la taxonomie des objectifs cognitifs de Bloom (voir § 2.2.2). En particulier, de nombreux praticiens croient que les questions fermées n'évaluent que les connaissances et non les capacités cognitives. Ils confondent en fait l'usage qui est généralement fait de ce type de questions et les possibilités effectives offertes par celles-ci. En réalité, tous les niveaux de capacité cognitive peuvent être évalués avec des questions fermées. De ce point de vue, les questions à choix multiple et les questions d'appariement offrent un potentiel rarement exploité. Les deux exemples suivants illustrent cette possibilité d'évaluer des capacités de haut niveau au moyen de questions fermées (d'après Wiersma & Jurs, 1990, p.53) :

1. Si a et b sont des nombres entiers et que a est plus petit que b, le rapport (a+5)/(b+5) est toujours :

   A. égal à un.

   B. plus grand que un.

   C. plus petit que un.
   D. un nombre négatif.
2. Lequel de ces processus ressemble le plus à la transformation de la glace en eau ?
   A. la dissolution d'un cube de sel dans l'eau.
   B. la fusion du minerai de fer dans un haut fourneau.
   C. la combustion du bois en fumée et en cendres.
   D. l'inspiration de l'oxygène et l'expiration du dioxyde de carbone.

Comme on peut le voir, les possibilités offertes par les questions fermées sont plus larges qu'on ne le pense habituellement. Leurs limites sont celles de l'imagination de leur créateur. En fait, ce que mesurent les questions fermées est déterminé plus par leur contenu que par leur format. Toutefois, il faut reconnaître que, par leur nature, certaines capacités ne peuvent être mesurées par des questions fermées. Il est évident que les capacités dactylographiques d'une secrétaire ne peuvent être évaluées qu'au travers d'un travail de dactylographie. De même, pour apprécier les capacités de rédaction d'un étudiant, il conviendra de lui demander de produire un texte écrit. D'une manière générale, lorsque l'évaluation veut prendre en compte la structuration et l'expression de la pensée, l'usage de questions ouvertes est nécessaire.

(2) Les conditions matérielles de l'évaluation

Les contraintes matérielles, tant au niveau de la préparation du test que de son administration, doivent également être prises en compte lors du choix du format des questions. Ces contraintes concernent le temps, l'espace et le matériel. Le temps de préparation des questions fermées est généralement beaucoup plus long que celui des questions ouvertes. En effet, la présentation aux sujets de plusieurs possibilités de réponses demande un travail de conception particulièrement délicat. Ce problème sera abordé plus en détail dans les § 3.1.3 à 3.1.5. Par contre, le temps de mise au point des questions fermées est souvent compensé par la brièveté du temps de cotation. Il suffit en effet de comparer les codes correspondant aux choix du sujet à ceux d'un tableau de référence. De plus en plus, les systèmes de lecture optique de protocoles permettent d'automatiser cette tâche. Outre leur vitesse (une centaine de protocoles peuvent être lus en quelques minutes), ces systèmes réduisent considérablement les risques d'erreur de codage et de transcription des résultats. Les codes lus par la machine sont enregistrés dans une base de données à partir de laquelle des calculs de scores et des grilles de résultats peuvent être produits très facilement.

Au contraire, les questions ouvertes prennent un temps de correction nettement plus long. C'est particulièrement le cas des questions demandant une réponse narrative. Ces dernières ont également comme inconvénient de demander beaucoup de temps au moment de la passation. Comme le font remarquer Ebel et Frisbie (1991), dans certains cas, les sujets passent plus de temps à rédiger leur réponse qu'à réfléchir au problème posé. Le temps de production des réponses narratives a pour conséquence de limiter l'étendue des connaissances évaluables en une seule séance. Il est alors nécessaire de prévoir plusieurs moments d'évaluation, ce qui n'est pas toujours possible.

Enfin, certaines contraintes matérielles doivent retenir l'attention du constructeur de test. Dans le cadre des évaluations scolaires ou des examens de recrutement, le

test doit souvent être passé par le sujet seul dans l'espace d'une classe ou d'un bureau. Des questions demandant des interventions répétées de l'examinateur (p.e. pour présenter du matériel ou pour poser des questions complémentaires) doivent alors être évitées. De même, le déplacement hors du local d'examen, la manipulation d'objets divers (p.e. : dictionnaire, pièces de puzzle...) est alors difficilement réalisable.

(3) LES FONCTIONS ASSIGNÉES AU TEST

L'usage qui sera fait du test pèse aussi lourdement sur le choix du format des questions. Les tests utilisés pour la certification ou la sélection doivent, le plus souvent, prendre en compte d'importantes contraintes de temps de passation et de correction. De plus, ces tests doivent avoir une fiabilité particulièrement élevée. En effet, ils débouchent généralement sur une décision, sans que d'autres prises d'information soient possibles. La mesure doit donc être très précise. Pour la même raison, ces tests doivent couvrir une étendue suffisante du domaine de compétence visé. Ces différentes contraintes font que des questions fermées sont généralement choisies pour ce type de test. Leur temps de passation et de correction est court, ce qui permet de poser de nombreuses questions couvrant largement le domaine visé. De même, leur fiabilité est bien contrôlée du fait de la standardisation des modalités de passation et de l'objectivité de la correction.

Les contraintes des tests diagnostiques et formatifs sont différentes. Le temps est moins contraignant. De plus, les prises d'information peuvent être régulières ce qui diminue les exigences de fiabilité et d'étendue du domaine couvert par les questions. Si cette couverture est trop étroite ou si l'erreur de mesure est trop importante, une évaluation ultérieure permettra souvent de corriger l'appréciation portée sur le sujet. C'est ce qui se passe régulièrement en situation de classe. Un échec lors d'une évaluation mal construite (question ambiguë, critères de correction inadéquats...) peut être nuancé par les évaluations suivantes. L'usage de questions ouvertes est souvent préféré dans les tests diagnostiques ou formatifs car elles ont la réputation de permettre un recueil d'information plus riche et plus approfondi à propos des compétences des sujets. Cet a priori doit toutefois être nuancé. Les questions fermées, en particulier les items à choix multiple, peuvent elles aussi fournir des informations diagnostiques très intéressantes. Si les distracteurs ont été choisis avec soin, une analyse des erreurs peut être réalisée sur l'ensemble du test. Par ailleurs, la validité des questions ouvertes ne doit pas être envisagée indépendamment de leur fiabilité. Si les résultats d'une question à réponse narrative sont entachés par une importante erreur, cela signifie que l'épreuve a mesuré autre chose que ce qui était visé. Autrement dit, sa validité est ipso facto affaiblie. La subjectivité de la correction est ici en cause. Trop souvent, les correcteurs n'ont pas de critère de correction suffisamment précis. Ils sont alors facilement influencés par des aspects de surface de la réponse (propreté, lisibilité, style d'écriture...) non pertinents pour les objectifs du test. Les sujets risquent également de bluffer dans les questions à réponse narrative. Ils masquent alors leur ignorance de l'essentiel en développant exagérément certains points de détail qu'ils connaissent relativement bien. Leur réponse est alors sensée mais non pertinente.

Dans certains cas, la modalité de réponse peut avoir une valeur formative. Proposer aux élèves de rédiger leurs réponses les obligent à structurer leur pensée et à

exprimer leurs idées dans une forme linguistiquement correcte. En ce sens, l'usage de questions ouvertes peut avoir une valeur pédagogique.

(4) Les risques liés à la suggestion de réponses

Un des problèmes essentiels des questions fermées est de suggérer des réponses. Cette suggestion peut avoir des conséquences indésirables qui doivent bien être évaluées par le constructeur d'un test. La plus importante est le risque de répondre au hasard *(guessing)*. S'il s'agit d'une question « vrai-faux », le sujet a une chance sur deux de répondre correctement de cette manière. S'il s'agit d'une réponse à choix multiple, la probabilité variera suivent le nombre d'alternatives proposées. Pour cette raison, les questions à choix multiple sont souvent préférées aux questions « vrai-faux ». L'impact du hasard est alors plus limité. À condition que les distracteurs soient également plausibles, la probabilité de réussir une question à choix multiple comprenant quatre alternatives n'est que de 1/4. Une façon de réduire l'impact du hasard est de pénaliser les erreurs. On accordera, par exemple, 2 points pour une réponse correcte, 0 point pour une réponse omise mais on retirera 1 point si la réponse choisie est erronée. Cette manière de coter, si elle est annoncée aux sujets, conduit ceux-ci à préférer l'omission plutôt que le choix de réponse au hasard. Une autre manière d'éviter le risque de réussite par chance est de recourir à des questions ouvertes à réponse brève. Par exemple, au lieu de demander de choisir entre quatre réponses possibles à un problème mathématique, un espace blanc peut être laissé pour inscrire la réponse. Dans ce cas, la question ouverte est aussi objective que la question fermée mais l'influence du hasard est considérablement réduite. De plus, la validité apparente (voir chapitre 5) est meilleure. En effet, les sujets ont souvent une perception plus positive de la validité d'une question ouverte que d'une question fermée, même si les deux évaluent une capacité identique.

L'impact du choix aléatoire des réponses ne doit toutefois pas être surestimé. En effet, on observe fréquemment que les sujets les plus faibles obtiennent des résultats inférieurs à ceux qu'ils auraient pu obtenir en choisissant leurs réponses au hasard. En d'autre terme, la stratégie du choix aléatoire n'est pas appliquée systématiquement par les sujets faibles. Au contraire, ceux-ci tentent malgré tout de répondre en s'appuyant sur certains indices de surface et tombent ainsi dans les pièges tendus par le constructeur du test.

Un dernier problème lié à la présentation des réponses est de suggérer des solutions fausses. Le sujet risque ainsi de mémoriser une réponse erronée. Ce problème a fait l'objet de nombreuses recherches qui relèvent l'importance de ce risque en début d'apprentissage (Leclercq, 1986). L'élève dont les connaissances sont en construction est en effet plus susceptible de retenir une réponse fausse qu'un élève dont les connaissances sont déjà bien structurées. La présentation des réponses risque également de surévaluer certains sujets, particulièrement si les questions portent sur des connaissances. En effet, un sujet dont l'apprentissage est inachevé et encore mal structuré peut être incapable de produire une réponse correcte alors qu'il peut reconnaître celle-ci parmi des distracteurs. Ce risque de surévaluation peut cependant être réduit en fonction de la qualité des distracteurs. Le premier exemple ci-dessous comprend des distracteurs qui peuvent être éliminés facilement par un sujet qui possède des connaissances historiques superficielle. Ces distracteurs sont en effet des réponses très

peu plausibles. Par contre, dans le second exemple, une plus grande maîtrise des connaissances est nécessaire pour pouvoir choisir la réponse correcte.

1. La période du règne personnel de Louis XIV s'étend de :
   A. 1814 à 1830.
   B. 1661 à 1715.
   C. 1515 à 1545.
   D. 1789 à 1804.
2. La période du règne personnel de Louis XIV s'étend de :
   A. 1661 à 1705.
   B. 1661 à 1715.
   C. 1638 à 1681.
   D. 1653 à 1715.

3.1.3 *Construire des questions à choix multiple*

Nous avons vu plus haut qu'une question à choix multiple est composée d'une amorce, qui pose le problème, suivie de plusieurs alternatives comprenant la solution correcte et des distracteurs. Une troisième composante de toute question à choix multiple n'avait pas encore été mentionnée : les consignes. Celles-ci décrivent la tâche demandée, la modalité de réponse et les règles de cotation. Une grande attention doit être accordée à la rédaction des consignes. En effet, tous les sujets ne sont pas familiers avec le format « choix multiple ». Il est donc nécessaire d'expliciter ce qui est attendu d'eux et comment ils doivent répondre. Même avec des sujets habitués à ce type de format d'item, il est utile de préciser clairement comment répondre. De nombreux problèmes sont ainsi évités au moment de la cotation (p.e. plusieurs réponses choisies, réponses fausses indiquées au lieu de la réponse correcte...). Enfin, les informations données à propos des principes de notation des réponses font partie d'une relation claire et honnête avec les sujets. Elles permettent à ceux-ci d'ajuster leur comportement en fonction de ce qui est attendu d'eux. Ceci est particulièrement important lorsque, par l'attribution d'une note négative aux réponses fausses, on veut éviter que les sujets ne répondent au hasard (voir §3.1.2).

La rédaction des questions à choix multiple de bonne qualité est une tâche complexe qui demande une excellente connaissance du domaine visé et des techniques de construction d'items. Pour rédiger une bonne question à choix multiple, quelques **règles de base** devraient être respectées :

(1) Avoir les idées claires à propos des connaissances et des capacités cognitives qui doivent être évaluées par les questions. De nombreuses questions sont mal rédigées simplement parce que leurs auteurs ne savent pas vraiment ce qu'ils veulent mesurer. Ils tendent alors à produire des items demandant un simple rappel de connaissances. Ce sont en effet les questions à choix multiple les plus faciles à construire.

(2) Clarifier au maximum la question en séparant nettement les informations à utiliser (par exemple, un texte documentaire ou les données d'un problème mathématique) et l'amorce.

EXEMPLE

« Lorsque nous regardons le monde dans sa globalité, il est clair que le problème du développement économique est le plus important ».

Cette phrase doit-elle être considérée comme :

A. un jugement de valeur.
B. une conclusion scientifique.
C. un fait établi.
D. une analogie.

Par ailleurs, plutôt que de répéter certaines informations dans les alternatives, il vaut mieux les regrouper dans l'amorce. Les deux exemples suivants illustrent la clarification qui peut être apportée en rassemblant plusieurs informations dans l'amorce.

EXEMPLE

1. Christophe Colomb :
   A. atteint le Nouveau-Monde à la recherche de richesses.
   B. voulait établir une colonie sur les côtes de l'Amérique du Sud.
   C. navigua jusqu'au Nouveau-Monde pour fuir les persécutions religieuses.
   D. espérait atteindre les côtes de l'Orient par l'est.
2. Le principal objectif du voyage de Christophe Colomb vers le Nouveau-Monde était :
   A. la recherche de richesses.
   B. l'établissement d'une colonie en Amérique du Sud.
   C. la fuite des persécutions religieuses.
   D. l'atteinte des côtes de l'Orient.

(3) Le choix des distracteurs est un problème crucial. Ceux-ci doivent être suffisamment vraisemblables sans quoi les sujets risquent de trouver les réponses correctes par simple élimination des alternatives invraisemblables (voir ci-dessus la question concernant le règne de Louis XIV). Une manière de procéder consiste à repérer les erreurs habituelles des élèves dans le domaine concerné. Dans l'exemple suivant, le choix *a* est une erreur d'opération (multiplication au lieu de division) ; le choix *b* est une inversion du mauvais nombre et le choix *c* est également une erreur d'opération (addition au lieu de division).

EXEMPLE

(1/4) / (2/3) =

A. 1/6
B. 8/3
C. 3/8
D. 11/12

Les alternatives peuvent aussi être des choix naturels. Par exemple, en néerlandais, un substantif peut être masculin, féminin ou neutre. Ces trois genres constitueront des alternatives naturelles dans une question portant sur le genre de substantifs néerlandais. De même, « présent, imparfait, futur » représentent des alternatives naturelles pour des questions portant sur le temps des verbes. Une autre manière de procéder pour trouver des distracteurs plausibles est de réflé-

chir aux éléments appartenant à la même catégorie que la réponse correcte (p.e. des animaux appartenant à la catégorie des félins si la réponse correcte est « chat ») ou qui sont naturellement associés à cette réponse (p.e. « bougie », « batterie »... si la réponse correcte est « ampoule »).

Dans la rédaction des alternatives, il y a lieu d'éviter les termes vagues (p.e. « parfois », « certain », « un peu »...) et les formulations négatives. Ils sont une source d'ambiguïté et de complexité. Ils risquent d'affaiblir la validité de la question. Par exemple, d'un sujet à l'autre, le terme « parfois » est associé à une fréquence d'événements très variable. La réponse choisie peut, par conséquent, différer en fonction de l'interprétation donnée à ce terme.

Quelques alternatives non classiques sont parfois utilisées dans les questions à choix multiple : « aucune des propositions », « toutes les propositions », « les propositions A et C »... Elles doivent être employées avec précaution. Certains praticiens les utilisent à mauvais escient lorsqu'ils ne trouvent pas d'autres alternatives. Les sujets repèrent vite un tel procédé et tendent à éliminer d'office cette alternative. Toutefois, bien utilisées, ces questions permettent de recueillir des informations intéressantes sur la qualité des apprentissages (voir Leclercq, 1986, pour une discussion détaillée).

Par ailleurs, quelques **erreurs fréquentes** doivent être évitées lors de la rédaction d'une question à choix multiple. Certains sujets peuvent en effet développer une véritable capacité *(test wiseness)* à utiliser ces vices de construction des questions pour repérer la réponse correcte parmi les distracteurs. Ils parviennent ainsi à obtenir des scores parfois élevés à des tests portant sur des domaines dont ils n'ont aucune connaissance. Les erreurs de construction les plus courantes sont :

(1) L'indication de la réponse correcte par une caractéristique grammaticale. Le pluriel et le genre des articles sont des indices fréquents. Ces indices peuvent être éliminés assez aisément en reformulant la question.

EXEMPLE

La tarentule est une :

A. mammifère.

B. reptile.

C. poisson.

D. araignée.

(2) La différence de longueur et de complexité des alternatives constitue un indice facile à repérer par les sujets intelligents. La solution est de construire des distracteurs dont la forme est plus proche de la réponse correcte.

EXEMPLE

Chez les scorpions, la fécondation se fait :

A. par contact.

B. de manière indirecte par l'intermédiaire du spermatophore.

C. par pénis et coït.

D. par les pattes.

(3) La répétition d'un même terme (ou partie de celui-ci) dans l'amorce et la réponse correcte est également un indice. Certains sujets répondent alors sur base des seules associations verbales. Une solution à ce problème est parfois difficile à trouver avec un format fermé. Par contre, une réponse ouverte brève permet aisément d'éliminer l'indice verbal.

EXEMPLE

Les arabes ont particulièrement développé un genre d'ornement :

A. les palmettes.
B. les feuilles d'acanthes.
C. les arabesques.
D. les fleurs de lotus.

3.1.4 *Construire des questions « vrai-faux »*

Nous avons vu dans le § 3.1.2 que les questions « vrai-faux » sont des propositions dont le sujet est invité à évaluer la véracité. Ces questions sont plus simples à créer que les questions à choix multiple puisque le problème de la construction de distracteurs vraisemblables est éliminé. Plus exactement, ce problème se retrouve uniquement dans la production des propositions fausses. Pour être efficaces, celles-ci ne peuvent être écartées sur base du seul bon sens ou d'indices de surface. Le jugement concernant leur fausseté doit nécessiter une réelle connaissance de la matière.

Avec des questions « vrai-faux », il est plus difficile d'évaluer des capacités cognitives de haut niveau. Toutefois, la complexité des opérations cognitives que ces questions permettent d'apprécier est souvent plus élevée qu'on ne le pense. Trop de praticiens se contentent de créer des questions qui ne demandent que le rappel de connaissances stéréotypées. Dans les pires cas, ces connaissances ne concernent que des détails peu importants. De telles questions offrent une image morcelée et anecdotique du savoir. Pourtant, bien construites, ces questions permettent d'apprécier si un élève a réellement compris les connaissances essentielles qui lui ont été enseignées. Les trois questions suivantes permettent d'évaluer différents niveaux de connaissance du principe d'Archimède (d'après Ebel & Frisbie, 1991). La première proposition demande le seul rappel d'une connaissance livresque. La seconde suppose une capacité de reformuler le principe étudié. Enfin, la troisième proposition fait appel à la capacité d'appliquer les connaissances apprises.

EXEMPLE

1. Un corps plongé dans un liquide subit une poussée verticale de bas en haut égale au poids du liquide déplacé.

   Vrai — Faux

2. Si un objet possédant un certain volume est entouré d'un liquide ou d'un gaz, la force de bas en haut qui s'exerce sur lui est égale au poids du même volume de liquide ou de gaz.

   Vrai — Faux

3. Lorsqu'ils sont immergés dans l'eau, un centimètre cube d'aluminium et un centimètre cube de fer subissent une même force de bas en haut.

   Vrai — Faux

Une des difficultés de la construction de questions « vrai-faux » est que le jugement les concernant doit être tranché. La proposition est soit vraie, soit fausse. Sa véracité ne peut être l'objet de variation ou de discussion. Ce problème est important car il sous-tend l'équité et la légitimité de l'évaluation qui sera faite à l'aide de ces questions. Il est en effet inacceptable de mesurer un degré de compétence à partir de jugements qui ne sont en réalité que des opinions. De même, il est fondé de contester les résultats d'un test composé de questions dont les réponses correctes ne sont pas défendables. Les deux exemples suivants sont des illustrations de propositions inadéquates pour un format « vrai-faux ».

EXEMPLE

1. Le poids d'un nuage de pluie est léger.

Vrai — Faux

2. Le mérite est un facteur important influençant le salaire des employés.

Vrai — Faux

Dans la formulation des questions « vrai-faux », il faut généralement éviter des déterminants comme « tous », « toujours », « aucun » ou « jamais ». Lorsque ces déterminants sont utilisés, la réponse correcte à la question est habituellement « faux ». En effet, il est rare qu'une affirmation ne souffre d'aucune exception. Il est en effet vraisemblable que, dans un cas au moins, l'affirmation est fausse. Dans le premier exemple ci-dessous, on ne peut exclure qu'un guerrier sioux ait manqué de courage. Par conséquent, la proposition doit être considérée comme fausse. Par contre, le second exemple est un cas, peu fréquent, où l'usage de « tous » est indiqué.

EXEMPLE

1. Tous les sioux étaient des guerriers courageux.

   Vrai — Faux

2. Tous les hydrates de carbones contiennent de l'oxygène, du carbone et des atomes d'hydrogène.

   Vrai — Faux

Les négations sont souvent une source de confusion dans les questions « vrai-faux », surtout lorsque le choix est entre « oui » et « non ». Si, par exemple, la proposition est *« il ne faut pas dépasser la vitesse de 60 km/heure en ville »*, le sujet peut entourer la réponse « non » parce qu'il pense que *« non, il ne faut effectivement pas dépasser la vitesse de 60 km/heure en ville »*. L'alternative « vrai-faux » réduit un tel risque de confusion, sans pour autant le faire disparaître. Par conséquent, il est préférable de toujours formuler les questions de manière affirmative.

Dans l'ensemble d'un test, il est préférable d'avoir un peu plus de propositions fausses que de propositions vraies. Les propositions fausses permettent de mieux discriminer les sujets faibles des sujets forts (Barker & Ebel, 1981) que les propositions vraies. En effet, en cas de doute, les sujets sont plus enclins à accepter les propositions présentées qu'à les refuser. Cette inclination est appelée *la tendance à l'acquiescement.* Par conséquent, au lieu d'inclure dans le test un même nombre de propositions vraies et fausses, comme le recommandent certains auteurs (p.e. Wiersa & Jurs, 1990), il vaut mieux respecter un rapport de deux propositions fausses pour une proposition

vraie afin d'obtenir un score total au test qui soit plus discriminatif (Ebel & Frisbie, 1991).

3.1.5 *Construire des questions d'appariement*

Ce format de question est utilisé moins couramment que les deux précédents. Rappelons qu'il se présente sous forme de deux colonnes. La première comprend les prémisses et la seconde les réponses. Les réponses doivent être associées à chacune des prémisses. Ce format a l'avantage de permettre l'évaluation de nombreuses connaissances en une seule question. Outre ce caractère économique, les questions d'appariement ont également l'avantage de ne pas nécessiter la création de distracteurs. Par contre, la cotation est un peu plus complexe. En effet, elle ne se fait pas au niveau de l'ensemble de la question mais pour chaque appariement. Par conséquent, si une question demande de réaliser quatre appariements, il faudra attribuer quatre scores aux réponses à cette question.

Une bonne question d'appariement doit être homogène. Si le contenu est trop hétérogène, les sujets risquent de trouver des indices leur permettant de répondre correctement tout en ayant très peu de connaissance du domaine évalué. L'exemple suivant est un cas de question trop hétérogène à laquelle il est possible de répondre avec un peu de bon sens, sans aucune connaissance spécifique.

EXEMPLE

....	1. Ville de la province du Hainaut	A.	la Meuse
....	2. Fleuve traversant la province de Liège	B.	Félicien Rops
....	3. Artiste célèbre de la province de Namur	C.	le bois
....	4. Industrie de la province du Luxembourg	D.	Binche

Pour éviter ce problème, une question d'appariement doit avoir un contenu homogène, c'est-à-dire se référant à un seul concept ou à une seule classe. On peut comparer de ce point de vue l'exemple précédent à celui qui suit.

EXEMPLE

La première colonne contient une liste de provinces belges. La seconde colonne contient, elle, une série de noms de villes. Associez chacune de ces villes avec la province à laquelle elle appartient.

....	1. Province du Hainaut	A.	Huy
....	2. Province de Liège	B.	Binche
....	3. Province de Namur	C.	Neuchâteau
....	4. Province du Luxembourg	D.	Dinant

Les deux exemples précédents proposent autant de réponses qu'il y a de prémisses. Une telle correspondance est déconseillée. En effet, il suffit que le sujet connaisse trois réponses correctes pour trouver automatiquement la quatrième. Pour éviter ce problème, il est recommandé de construire des questions asymétriques : soit par excès de prémisses, soit par excès de réponses. Il est également possible d'utiliser une même réponse pour plusieurs prémisses.

EXEMPLE

La première colonne contient une liste d'événements qui se produisent dans la vie quotidienne. La seconde colonne contient, elle, une série de termes scientifiques qui décrivent ces événements. Indiquez devant chaque événement le terme qui lui correspond

1.	La glace fond	A. L'expansion
.... 2.	Les vêtements sèchent	B. La condensation
.... 3.	Les nuages se forment	C. La fusion
.... 4.	La pluie tombe	D. L'évaporation
		E. La précipitation
		F. La radiation

3.1.6 *Construire des questions ouvertes*

Les réponses ouvertes sont souvent choisies pour des raisons de validité. Dans certains cas, il est évident que les questions ouvertes mesurent mieux certaines compétences que ne le font les questions fermées. De plus les sujets ont généralement une meilleure perception des questions ouvertes qui leur paraissent plus légitimes et avec lesquelles ils ont le sentiment de mieux maîtriser la situation. Nous avons déjà vu plus haut que trois formes de questions ouvertes peuvent être utilisées : (1) les questions à réponse brève, (2) les réponses narratives et (3) les réponses qui demandent la production d'un comportement. Dans cette section nous allons détailler les deux premières formes de questions ouvertes.

Les *questions à réponse brève* sont assez proches des questions fermées. Elles ne sont dites ouvertes que parce que la réponse n'est pas donnée et que le sujet doit donc la produire. Mais leur cotation peut être aussi objective que celle des questions fermées car une seule réponse correcte est possible. L'intervention de la subjectivité du correcteur est alors nulle. Ceci est vrai pour autant que la question ait été bien construite. Il n'est pas toujours simple de créer des questions dont la réponse est unique et tient en un seul mot ou en un seul nombre. Pour parvenir à un tel résultat, il est préférable de commencer par penser à la réponse puis d'élaborer une question qui doit déboucher sur cette réponse. Pour éviter toute variabilité au moment de la correction, il est également important de préciser dans la question certaines caractéristiques qui doivent être présentes dans la réponse. Ainsi, lorsque la réponse attendue est numérique, il faudra annoncer dans la question la précision du résultat attendu (nombre de chiffres après la virgule) et l'éventuelle unité de mesure (centimètre, litre...) à mentionner dans la réponse. Plusieurs erreurs de construction doivent également être évitées. Il ne faut pas donner des indices à propos de la réponse correcte dans la formulation de la question. En particulier, les espaces prévus pour noter les réponses doivent être de même longueur pour toutes les questions d'un même test. Il est fréquent que les sujets infèrent la réponse correcte en se basant sur l'espace laissé pour répondre.

Parfois, plusieurs réponses courtes sont regroupées sous un même problème. Par exemple, il peut s'agir de la description des symptômes d'un malade suivie de questions relatives au diagnostic, au pronostic et au traitement. Les questions peuvent être totalement indépendantes les unes des autres. Dans l'exemple, elles sont au contraire liées. L'étudiant qui échoue à la première question ne peut donner les réponses attendues aux questions suivantes. S'il ne pose pas un diagnostic correct, il ne peut en

effet proposer le traitement adapté. Généralement, de telles questions emboîtées doivent être évitées car elles défavorisent indûment les sujets qui échouent les premières questions, ce qui n'est pas le cas pour les sujets qui échouent les dernières questions. Toutefois, dans certains cas, les questions emboîtées sont tout à fait indiquées. L'exemple ci-dessus en est une bonne illustration. L'étudiant en médecine à qui l'on présente un tel ensemble de questions doit nécessairement réussir toutes celles-ci. En effet, il n'est pas admissible qu'un médecin ne réponde que partiellement aux problèmes qui se posent à lui : le patient est soigné correctement ou non. Il serait discutable d'accorder des points à un étudiant qui propose un traitement correct pour soigner une maladie dont le diagnostic est erroné.

Les *questions demandant une réponse narrative* apparaissent, quant à elles, comme le prototype des réponses ouvertes. Elles sont bien adaptées pour évaluer des compétences de haut niveau comme la résolution de problèmes complexes, l'intégration des connaissances, l'esprit critique et la créativité. Ces questions sont souvent perçues comme plus faciles à construire que les autres formats de question. En fait, quelques règles devraient être respectées pour leur création si l'on veut éviter certains déboires au moment de la cotation. En particulier, il est nécessaire de donner aux sujets des informations précises et complètes à propos de ce qui est attendu d'eux. De nombreux problèmes surgissent au moment de la correction de réponses narratives simplement parce que les sujets ont interprété différemment la question posée. Dans ce cas, qui faut-il blâmer : le constructeur de la question ou le sujet examiné ? Comme souvent le constructeur est le correcteur, la faute est évidemment reportée sur le sujet (« il n'a rien compris à ce qu'on lui demandait ! »). Les termes utilisés dans la rédaction des questions devraient toujours faire référence aux capacités cognitives que l'on souhaite évaluer : « expliquer... », « comparer... », « interpréter... », « critiquer... », « évaluer... ». Si l'on désire limiter les réponses à une certaine longueur ou obliger les sujets à respecter une certaine structure, il est possible de proposer des *questions à réponse contrainte* (Gronlund, 1991, p.76). Dans ce cas, la question contient un certain nombre de directives concernant la forme de la question. Par contre, les *questions à réponse développée* laissent toute liberté aux sujets quant à la longueur et à la structuration de leur réponse. Une telle latitude permet plus de créativité et une approche plus large du problème posé mais elle est source de complexité au moment de la correction.

EXEMPLE DE QUESTIONS À RÉPONSE CONTRAINTE

1. Expliquez en une demi-page, les avantages des questions ouvertes.
2. Un professeur de science veut, au moyen d'un test papier-crayon, évaluer les aptitudes de ses élèves à interpréter des données scientifiques.
   – Décrivez les étapes que devrait suivre ce professeur.
   – Donnez des arguments pour justifier chacune de ces étapes.

EXEMPLE DE QUESTION À RÉPONSE DÉVELOPPÉE

Vous êtes professeur de science. Planifiez de manière complète une évaluation sommative des acquisitions de vos élèves. Détaillez chacune des procédures que vous pensez suivre, les instruments que vous souhaitez utiliser et les raisons de vos différents choix.

3.2 Formats d'items pour les questionnaires

L'évaluation de traits de personnalité, d'attitudes, d'intérêts, de valeurs... fait appel à certains formats d'items particuliers. Dans le cas de la personnalité, des questions ouvertes demandant une performance sont souvent utilisées. Les *techniques projectives* en sont l'illustration la plus connue. Ces techniques consistent en un ou plusieurs stimuli (images, figurines, propositions...) à partir desquels le sujet est invité à produire des associations verbales, un récit, un dessin ou une construction. Ces productions sont considérées comme des manifestations de la structure profonde de la personnalité du sujet. L'information recueillie de la sorte est souvent riche mais difficile à coter. Des systèmes précis de cotation ont été mis au point pour certaines techniques projectives, en particulier pour le test de Rorschach (p.e. Exner, 1974). Ces systèmes demandent une bonne formation des correcteurs et leur application rigoureuse prend beaucoup de temps. Ils garantissent toutefois une fiabilité et une validité satisfaisantes des résultats, pour autant que les praticiens les respectent, ce qui n'est pas toujours le cas. Une étude faite par Exner et Exner (1972) auprès de 750 membres de la Society for Personnality Assessment et de l'American Psychological Association révèle en effet une grande diversité de pratiques de cotation du Rorschach. Vingt pour-cent des praticiens avouent ne faire aucune cotation objective et interpréter les réponses subjectivement sur base de leur expérience personnelle. Et quatre praticiens sur cinq reconnaissent personnaliser leur cotation. Par ailleurs, la majorité des autres techniques projectives reposent sur une standardisation insuffisante des modalités de passation et de cotation. Il en résulte des problèmes sérieux de fiabilité et de validité des résultats qu'elles permettent de recueillir (Klopfer & Taulbee, 1976).

Les questionnaires en auto-passation et les échelles de cotation (questionnaires remplis par un observateur et non par le sujet lui-même) sont nettement plus standardisés que les tests demandant une performance. Leur fiabilité et leur validité sont, par conséquent, plus assurées. Toutefois, nous verrons plus loin que les questionnaires présentent également certaines faiblesses spécifiques qui peuvent réduire la validité des résultats qu'ils permettent de recueillir. Trois formats d'items sont habituellement utilisés dans les questionnaires : les *items dichotomiques*, les *items catégoriels bipolaires* et les *items à choix forcé*. Nous allons en détailler les caractéristiques.

3.2.1 *Les items dichotomiques*

Un item dichotomique est constitué d'une proposition par rapport à laquelle le sujet doit exprimer son accord ou son désaccord. Le choix peut être entre « d'accord-pas d'accord », « oui-non, « vrai-faux »....

Exemple:

1	J'ai peu d'appétit	OUI - NON
2.	J'aime parler avec les personnes de mon entourage	OUI - NON
3.	Je n'ai aucun projet	OUI - NON
4.	J'ai envie de mourir	OUI - NON

La construction d'items dichotomiques est, en apparence, assez simple. Ce format soulève pourtant plusieurs problèmes dont certains sont difficiles à résoudre. Le

premier problème tient à la formulation des propositions. Dans l'exemple ci-dessus, une des propositions est formulée de manière négative (*« je n'ai aucun projet »*). Cette formulation complexifie la tâche du répondant. Doit-il répondre *« NON, je n'ai aucun projet »* ou *« OUI, je n'ai aucun projet »* ? Le premier choix correspond à une formulation plus naturelle que celle qui correspond au second choix. Pourtant, le sujet qui est d'accord avec la proposition doit choisir *« OUI »*. L'utilisation des modalités de réponse « vrai-faux » ou « d 'accord-pas d'accord » peut réduire ce problème.

Un second problème est lié au caractère tranché du choix demandé au sujet. Si la formulation de l'item est trop vague, celui-ci peut hésiter à choisir une des alternatives. Par exemple, si la proposition est *« je suis une personne relativement inquiète »*, l'interprétation du terme *« relativement »* peut varier et entraîner un choix qui dépende de cette interprétation. Par conséquent, de tels termes doivent être évités afin que les choix proposés soient identifiés de manière claire et identique par tous les sujets.

Un troisième problème posé par les items dichotomiques découle du phénomène de *désirabilité sociale*. De nombreux sujets ont en effet tendance à masquer leur véritable choix et à sélectionner, au contraire, le choix opposé par ce qu'il est plus valorisé socialement. Cette tendance peut découler d'un refus des sujets de se voir tels qu'ils sont et/ou d'une crainte du regard que le psychologue peut porter sur eux. Des propositions telles que *« je suis grossier »*, *« je ne pense qu'à moi »* ou *« j'aime la violence »* risquent ainsi de faire l'objet d'un choix négatif même si elles correspondent effectivement aux caractéristiques des sujets concernés. Pour éviter ce biais dû à la désirabilité sociale, on cherche généralement à créer des propositions moins transparentes. Elles doivent être des indicateurs valides tout en étant acceptables socialement. Par ailleurs, certains questionnaires incluent des items spécialement destinés à repérer l'impact de la désirabilité sociale. Par exemple, le *Minnesota Multiphasic Personality Inventory* (MMPI) comprend 15 propositions que quasi tous les sujets admettent comme vraies, tout en les considérant peu flatteuses. Un item comme *« mes manières de table ne sont pas toujours aussi bonnes à la maison qu'elles le sont en compagnie »* peut raisonnablement être considéré comme peu flatteur tout en étant vrai pour tout le monde. Pourtant, certaines personnes répondent systématiquement *« NON »* à de telles propositions. Ces personnes démontrent ainsi l'influence qu'a la désirabilité sociale sur leurs choix. On peut en déduire que les réponses à l'ensemble des items du questionnaire ont été biaisées par ce facteur. Malheureusement, l'usage de tels items pour diagnostiquer l'impact de la désirabilité sociale n'est réellement efficace qu'avec des sujets peu subtils. La majorité des individus ne se laisse généralement pas abuser par ces items. De plus, le diagnostic est fait après coup. La validité des résultats au questionnaire est alors mise en question sans qu'il soit possible de l'améliorer. Nous verrons plus loin que les items à choix forcé peuvent constituer une solution à ce problème.

Un dernier problème posé par les items dichotomiques concerne le calcul du score global. Généralement, les réponses aux items sont cotées 1 ou 0. La valeur 1 indique la présence de la caractéristique mesurée et la valeur 0 son absence. En fonction des propositions, une réponse *« OUI »* peut donc être cotée 1 ou 0. Revenons à l'exemple des quatre propositions ci-dessus qui servent à évaluer la dépression. Répondre *« OUI »* à la première est un signe de dépression et doit donc être coté 1. Par contre, répondre *« OUI »* à la seconde proposition indique une humeur normale et doit

être coté 0. Le plus souvent, le résultat total est calculé en additionnant les scores aux différents items. Par conséquent, chaque item a un poids identique dans le score total. Cette façon de faire a l'avantage de la simplicité. Sa pertinence est toutefois discutable. Tous les items n'indiquent pas un même degré du trait mesuré. Par exemple, une réponse positive à l'item *« j'ai peu d'appétit »* n'indique pas une même intensité de dépression qu'une réponse positive à l'item *« j'ai envie de mourir »*. Une solution à ce problème est de pondérer les résultats des items et d'ainsi donner un poids plus grand aux items en fonction de l'intensité du trait qu'ils permettent de révéler. La procédure la plus ancienne pour pondérer les scores aux items a été proposée par Thurstone (1928). Cette procédure est appelée *la technique des intervalles approximativement égaux* (« equal-appearing interval technique »). Elle consiste à placer chaque item sur le continuum à mesurer et à définir une échelle approximativement d'intervalle, généralement appelée *« échelle de Thurstone »*. Bien qu'encore décrite dans plusieurs ouvrages récents (p.e. Dane, 1990), la procédure de Thurstone a surtout une valeur historique. Elle est avantageusement remplacée par les procédures développées dans le cadre des *Modèles de la Réponse à l'Item* (Hambleton & Swaminathan, 1985, p. 115-120 ; voir également le chapitre 8).

3.2.2 *Les items catégoriels bipolaires*

Face à certaines propositions, il est possible de donner des réponses plus nuancées que *« d'accord »* ou *« pas d'accord »*. Des catégories intermédiaires peuvent être définies entre ces deux pôles opposés. L'ensemble des choix constitue des catégories ordonnées. Celles-ci forment ce que l'on appelle une *échelle de Likert*. Le nombre de catégories peut varier mais se limite généralement à cinq, comme l'a d'ailleurs suggéré Likert (1932). Ces catégories sont : *« en total désaccord »*, *« pas d'accord »*, *« neutre »*, *« d'accord »*, *« en total accord »*. D'autres termes équivalents peuvent être utilisés. Chaque catégorie se voit attribuer respectivement les scores 0, 1, 2, 3 et 4.

EXEMPLE

« L'étude des statistiques est nécessaire à la formation du psychologue »

pas du tout d'accord ❑ *pas d'accord* ❑ *neutre* ❑ *d'accord* ❑ *tout à fait d'accord* ❑

« J'éprouve des difficultés à m'endormir »

jamais ❑ *rarement* ❑ *parfois* ❑ *souvent* ❑ *très souvent* ❑

Comme les items dichotomiques, les items catégoriels bipolaires sont sensibles à l'influence de la désirabilité sociale. Mais ils soulèvent aussi des problèmes spécifiques comme la tendance à donner une réponse centrale. Pour contrecarrer cette tendance, on peut choisir de limiter le nombre de catégories à quatre. Par ailleurs, les items catégoriels bipolaires sont plus complexes à construire que les items dichotomiques. La définition des différentes catégories de réponses n'est pas toujours simple. Leur nombre et leur gradation peuvent poser problème.

3.2.3 *Les items à choix forcé*

Ce format a été crée pour tenter de résoudre un des problèmes posé par les formats précédents : l'influence de la désirabilité sociale sur le choix de la réponse. Un item à choix forcé, appelé *tétrade*, est habituellement constitué de quatre éléments :

deux indicateurs valides d'un trait et deux indicateurs non valides de ce même trait. Un des deux indicateurs valides est désirable socialement alors que l'autre ne l'est pas. Il en va de même pour les deux indicateurs non valides. Les sujets sont invités à choisir dans la tétrade la caractéristique qui leur correspond le mieux et celle qui leur correspond le moins. Pour chaque item, les sujets doivent donc donner deux réponses. L'exemple suivant (Guilford, 1954, p.275) est une illustration d'un tel item :

	Me correspond le plus	Me correspond le moins
1. peu soigneux	❑	❑
2. sérieux	❑	❑
3. énergique	❑	❑
4. snob	❑	❑

La création de tels items se fait en plusieurs étapes. La première étape consiste à observer ou à interviewer des personnes qui possèdent le trait à mesurer à un degré très élevé ou très faible. Sur base de ce recueil d'informations, des indicateurs du trait sont produits. Il s'agit de qualificatifs, de substantifs ou de courtes phrases qui sont associés à l'absence ou à la présence du trait. La validité de ces termes est ensuite évaluée par des experts et par le calcul d'un index de validité (voir chapitre 6). Sur base de ces évaluations, des pairs d'indicateurs sont constituées. Toutes comprennent un indicateur valide et un indicateur non valide de même niveau de désirabilité. On veille à constituer des paires d'indicateurs également désirables et d'autres également indésirables. Une fois ces paires réalisées, on peut alors construire des tétrades en groupant chaque fois une paire d'éléments désirables et une paire d'éléments non désirables. Dans l'exemple ci-dessus, « sérieux » et « énergique » sont deux caractéristiques également désirables socialement. La première s'est révélée valide lors d'études préliminaires, mais pas la seconde. Quant à « peu soigneux » et « snob », il s'agit de caractéristiques également indésirables. La première est valide mais pas la seconde.

Un sujet qui a tendance à se présenter sous un jour trop favorable a autant de chance de choisir un indicateur valide qu'un indicateur non valide. Lorsqu'il choisit un qualificatif désirable mais non valide, ce choix n'influence pas le score total. En réalité, ce choix a pour effet de déprimer le score global. Il empêche en effet un autre choix, valide celui-là, qui aurait pu augmenter le score total. L'impact de la désirabilité est dès lors réduit. Le même phénomène se produit lorsqu'un sujet tend systématiquement à se dévaloriser.

4. Conclusion

Comme nous l'avons souligné au début de ce chapitre, la création des items est un moment crucial dans la construction d'un test. La qualité de ce travail détermine la valeur de l'instrument dans son ensemble. Pourtant, depuis plus de cinquante ans, les chercheurs ont concentré beaucoup plus leur attention sur l'étude des propriétés métriques des items que sur la méthodologie de leur construction. Par conséquent, la création des items reste le plus souvent basée sur l'intuition et le bon sens. Les praticiens comptent alors sur les analyses statistiques ultérieures pour débusquer les mauvais

items. S'ils ont de la chance, les items posséderont dans leur majorité les propriétés voulues et ils pourront rapidement passer à la phase suivante du travail de mise au point du test. Mais, souvent, les items faibles seront trop nombreux. Il sera alors nécessaire soit de créer un certain nombre de nouveaux items, soit de reconstruire l'ensemble des items selon de nouveaux principes. Cette situation est bien entendu coûteuse en temps et en énergie. Une économie substantielle aurait pu être réalisée en apportant plus de soin à la création de l'ensemble d'items initial. Dans le présent chapitre, nous avons indiqué quelques pistes méthodologiques permettant de garantir une certaine qualité des items. Il ne s'agit cependant pas de recettes miraculeuses. Tout chercheur qui a eu l'occasion de construire un test sait que des items apparemment bien construits peuvent réserver de mauvaises surprises sur le terrain. Mais un travail de création méthodique permet de limiter au maximum le nombre des items défectueux. Le prétest jouera, quant à lui, un rôle de contrôle de qualité en nous révélant les inévitables faiblesses de quelques items. Dans le chapitre 6, nous aborderons en détail les différentes techniques statistiques permettant d'évaluer les items et de repérer leurs éventuels défauts.

CHAPITRE 4

MODÈLES CLASSIQUES DES TESTS

Les théories des tests nous offrent un cadre conceptuel pour apprécier la valeur des résultats obtenus au moyen d'instruments de mesure. Chaque théorie s'appuie sur une conception particulière de la mesure et sur une série de postulats sur la nature des données et sur la manière dont elles ont été recueillies. Quels que soient les postulats, il est important de réaliser que chaque théorie n'est qu'un *modèle* simplifié de la réalité. Chaque modèle s'ajuste plus ou moins bien à la réalité qu'il cherche à décrire. C'est pourquoi une connaissance minimale de ces modèles est nécessaire pour pouvoir apprécier s'ils sont adaptés aux conditions de mesure rencontrées et s'ils permettent de répondre à nos besoins.

Les modèles parfaits n'existent pas. La tentation peut être forte d'*adapter les données au modèle* plutôt que d'*adapter le modèle aux données*. La première approche revient à se mentir à soi-même. Tout modèle étant une simplification de la réalité nous devons être conscients de ses limites et des implications de celles-ci pour l'interprétation des données.

1. Propriétés des scores composites

La théorie classique des scores a pour principal objet le score total obtenu par chaque personne à un test. Or, ce score total est un score composé de la somme des résultats à chaque item pris individuellement. Avant d'aborder les questions de fiabilité et de validité de ce score total, il est important de décrire comment le score total est lié aux items qui le composent. Ce qui nous préoccupe particulièrement, c'est de connaître comment la variance de ce score total se répartit en fonction des différents items. La variance totale des scores est importante en théorie classique puisque, pour différencier les personnes, il faut que les résultats possèdent une certaine variance. Il y a peu d'utilité à discuter de questions de fiabilité ou de validité s'il n'y a aucune différence entre les individus.

1.1 COMBIEN FONT DEUX ORANGES PLUS TROIS CITRONS ?

Le score total est un score composite : il est le résultat de l'addition des scores aux items du test. En effectuant cette addition, nous postulons que ces items mesurent sensiblement le même trait. Que signifie, par exemple, le score global d'un test composé de questions de géographie et de mathématiques ?

Selon le degré de pertinence que nous souhaitons retrouver dans les unités de mesure de notre score total, nous choisirons d'additionner des éléments provenant d'ensembles dont la définition en compréhension est très stricte ou, au contraire, relativement large. À la question « combien font deux oranges et trois citrons ? », nous pouvons répondre de trois manières différentes :

1. nous pouvons refuser de faire l'addition, considérant qu'il s'agit de deux catégories différentes ;
2. nous pouvons ramener chaque ensemble à un ensemble qui les contient tous les deux (p.e. la catégorie des fruits) et effectuer l'opération à l'intérieur de cet ensemble plus large. Dans notre exemple, la réponse est alors « cinq agrumes » ou « cinq fruits » ;
3. nous pouvons aussi ignorer les caractéristiques communes à chaque ensemble et répondre comme plusieurs jeunes enfants de six ans : « deux pommes plus trois oranges font cinq compotes ».

Chacune des solutions précédentes trouve un écho dans le calcul des scores à un test. La première solution consiste à calculer non pas un score total, mais un profil de performance. Le praticien refuse de confondre entre elles certaines réussites et préfère calculer des sous-scores. Cette procédure est particulièrement utilisée avec la mesure critériée.

La seconde solution consiste à faire abstraction des particularités de chaque ensemble pour ne prendre en compte que les caractéristiques générales. L'addition d'items différents est possible dans ce contexte en postulant qu'ils ont tous au moins quelque chose en commun. Ce quelque chose peut être plus ou moins vague. En additionnant deux citrons et trois oranges, on peut répondre « cinq agrumes » ou « cinq fruits ». La première des deux réponses est certainement la plus pertinente. Pour évaluer les apprentissages scolaires, plus les objectifs d'évaluation sont précis et bien hiérarchisés, plus grande sera la validité de contenu ou, si l'on préfère, la pertinence des résultats. Ce genre de préoccupation trouve sa place dans l'*évaluation sommative*. Lorsque le domaine à mesurer est vaste, il faut que les items puissent échantillonner une grande étendue de contenu. Il en résulte que pour obtenir un score total qui couvre une matière plus vaste, il faut faire abstraction de certaines caractéristiques des items.

La troisième solution reviendrait à additionner les résultats à des items sans savoir de façon précise ce que chacun mesure. C'est le danger que l'on court à additionner un méli-mélo d'items qui ont été rédigés sans cadre préalable. Les réponses « cinq végétaux » ou « cinq choses » n'ont de précis que le chiffre.

La théorie classique des scores ne traite que de la précision de la valeur numérique. La qualité ou la pertinence de cette valeur sont traitées séparément par l'intermédiaire d'études de validité. Face à un résultat de soixante pour cent, la fiabilité consiste

à se demander : « est-ce bien soixante ? », alors que la validité pose la question « soixante pour cent de quoi ? ». Pour déterminer ce que signifie un score total, il faudra donc dépasser la question de sa précision. En aucun cas, une grande assurance en la valeur numérique des résultats ne doit nous faire oublier la question importante de sa signification, que nous verrons dans le chapitre 5 consacré à la validité.

1.2 VARIANCE TOTALE DES RÉSULTATS À UN TEST

Si nous sommes intéressés à différencier entre eux des individus, alors la variance des résultats est une caractéristique importante. Rappelons que la variance d'une distribution de résultats est égale à la somme des écarts quadratiques (au carré) à la moyenne divisée par le nombre de résultats. C'est ce que traduit l'équation suivante :

$$S^2_x = \sum \frac{(X - \bar{X})^2}{n} \qquad (4.1)$$

Partant de cette formule, nous pouvons aisément nous rendre compte que la variance d'un ensemble de scores sera d'autant plus grande que plusieurs sujets obtiennent des résultats différents de la moyenne et réciproquement. Par exemple, un individu qui est à 10 points de la moyenne ajoute 100/n à la variance, alors qu'un individu qui est à 2 points de la moyenne n'ajoute que 4/n à la variance totale, soit 25 fois moins pour un écart à la moyenne 5 fois plus petit. En fait, un résultat extrême peut même faire paraître la variance totale des scores bien supérieure à ce qu'elle est en réalité. Le tableau 1 en présente un exemple concret.

Le tableau 1 présente cinq cas différents. Dans chaque cas, seule la première valeur est changée, celle du sujet #1. La situation initiale (cas # 1) est celle d'une distribution dont la moyenne est 84,73 et la variance de 58,93. Lorsque, comme dans le cas #2, l'on change la valeur du premier sujet de 71 à 81 (en direction de la moyenne du groupe), la variance totale diminue à 42,23. Par contre, si le même changement de 10 points s'effectue dans une direction opposée à celle la moyenne (cas #3), la variance totale passe de 58,93 à 92,15 et la somme des carrés des écarts de 648,18 à 1013,64 (presque le double). Lorsque ce changement n'est que de deux (cas #4), la variance passe de 58,93 à 54,26, une différence de 4,67, lorsqu'il s'effectue en direction de la moyenne. En guise de rappel, l'écart était de 16,70 pour un changement de 10 (cas #2). Le changement de variance est presque 3 fois plus grand lorsque le score du sujet #1 voit son écart à la moyenne passer de deux à 10. Enfin, le cas #5, illustre ce qui se produit lorsque l'écart de deux s'effectue en direction opposée à la moyenne. La variance passe de 58,93 à 64,25, une différence de 5,32. En comparaison avec le cas #3, il s'agit d'une différence bien moindre, alors que pour un écart de 10 opposé à la moyenne le changement de variance produit était de 33,22 (92,15 - 58,93). Le tableau 1 illustre à quel point une erreur même minime de codage des données peut avoir une grande répercussion sur la variance, alors que la moyenne est pour sa part beaucoup moins affectée par ces changements. En outre, le poids des résultats extrêmes dépend aussi de la taille des échantillons. Plus celui-ci est petit, plus l'impact sera grand sur la variance.

Tableau 1 – Cinq cas de variance totale

Sujet#	Situation initiale	Changement de 10 vers la moyenne	Changement de 10 opposé à la moyenne	Changement de 2 vers la moyenne	Changement de 2 opposé à la moyenne
	Cas 1	*Cas 2*	*Cas 3*	*Cas 3*	*Cas 4*
1	71	81	61	73	69
2	75	75	75	75	75
3	79	79	79	79	79
4	82	82	82	82	82
5	84	84	84	84	84
6	85	85	85	85	85
7	86	86	86	86	86
8	87	87	87	87	87
9	90	90	90	90	90
10	94	94	94	94	94
11	99	99	99	99	99
Somme	932	942	922	934	930
Moyenne	84,73	85,64	83,82	84,91	84,55
Carré des écarts	648,18	464,55	1013,64	596,91	706,73
Variance	58,93	42,23	92,15	54,26	64,25
Écart-type	7,68	6,5	9,6	7,37	8,02

Voyons à présent comment la variance est affectée par les résultats aux items d'un test. En effet, chaque item possède un impact particulier sur le score total à un test, sur sa moyenne et aussi sur sa variance.

Prenons un exemple fort simple. Supposons qu'un item soit réussi par tous les sujets. Cet item ne possède aucune variance. Son rôle dans le score total se réduit à accroître la moyenne. Par contre, il n'ajoute aucune information supplémentaire nous permettant de départager entre eux les personnes ayant répondu au test. La même situation prévaudrait dans le cas d'un item qui serait échoué par tous les sujets.

Le tableau 2 illustre cette situation. Considérons un test constitué de deux items (items 1 et 2) administré à sept (7) personnes. Ajoutons-y l'item 3 réussi par tous. Le résultat du nouveau test comprenant maintenant trois items (X+3) indique que la

moyenne s'est accrue de 1 (la moyenne de l'item 3 + la moyenne du score X constitué des deux premiers items). Quant à la variance totale du nouveau test, elle n'a pas changé. L'item 3 étant réussi de façon constante par tous, sa variance est nulle et par conséquent l'ajouter au score total ne change rien.

L'item 4 n'est pas réussi par tous. En fait, il possède une variance de 0,24. Si on ajoute son résultat à celui du score X, on remarque que là aussi la moyenne augmente. Comme précédemment, la nouvelle moyenne est le résultat de la somme de la moyenne des deux premiers items plus celle de l'item 4. À la différence de l'item 3, cependant, l'ajout de l'item 4 change également quelque chose à la variance totale du test. La variance du test X+4 est en effet de 2,20, alors qu'elle n'était que de 0,98 pour le test X. Notez bien que la variance totale du test X+4 est bien plus que la somme des variances du test X et de l'item 4. En effet, 0,98+0,24 < 2,20. On peut aussi remarquer que la variance totale du score X est bien supérieure à la somme des variances des items 1 et 2 (0,24+0,24 < 0,98). À quoi peut-on attribuer ces différences ?

Tableau 2 – Variance d'un score composite. Exemple 1

Sujet#	Item 1	Item 2	Score X	Item 3	Item 4	X+3	X+4
1	1	1	2	1	1	3	3
2	0	0	0	1	0	1	0
3	1	1	2	1	1	3	3
4	0	0	0	1	0	1	0
5	0	0	0	1	0	1	0
6	1	1	2	1	1	3	3
7	1	1	2	1	1	3	3
Moyenne	0,57	0,57	1,14	1	0,57	2,14	1,71
Variance	0,24	0,24	0,98	0	0,24	0,98	2,2

Matrice des variances-covariances

	Item 1	Item 2	Item 3	Item 4
Item 1	0,24	0,24	0	0,24
Item 2	0,24	0,24	0	0,24
Item 3	0	0	0	0
Item 4	0,24	0,24	0	0,24

Voyons l'exemple du tableau 3. Dans ce deuxième exemple, la somme des variances des items 1 et 2 est supérieure à celle du score total X (0,24+0,20 > 0,20). Comment cela est-il possible ? Un tel résultat nous amènerait à conclure qu'il est plus facile de différencier les personnes à partir du résultat à un seul item au test qu'à partir

des résultats au test entier. Pourquoi cette différence entre ce que nous observons au tableau 2 et au tableau 3 ?

La réponse se situe dans l'examen attentif de la relation qui existe entre les deux items constituant chacun des tests. Dans le tableau 2, les items sont réussis ou échoués simultanément. Dans le tableau 3, l'information fournie par les items est plus contradictoire : lorsqu'un item est réussi, l'autre est souvent échoué et vice versa. Nous pouvons dire que les résultats au premier test sont *homogènes*, alors que les résultats au second test sont *hétérogènes*.

Lorsque nous devons constituer un score total, il est préférable d'additionner ensemble des items homogènes plutôt que des items hétérogènes. En fait, il n'y pas d'intérêt à additionner des items hétérogènes lorsque notre objectif est de différencier des individus. Leurs valeurs différentes ont tendance à s'annuler ce qui a pour effet de réduire la variance du score total au test.

Tableau 3 – Variance d'un score composite. Exemple 2

Sujet#	Item 1	Item 2	Score X	Item 3	Item 4	X+3	X+4
1	1	0	1	1	1	2	2
2	0	1	1	1	1	2	2
3	1	0	1	1	1	2	2
4	0	0	0	1	0	1	0
5	0	0	0	1	0	1	0
6	0	1	1	1	1	2	2
7	1	0	1	1	1	2	2
Moyenne	0,43	0,29	0,71	1	0,71	1,71	1,43
Variance	0,24	0,2	0,2	0	0,2	0,2	0,82

Matrice des variances-covariances

	Item 1	Item 2	Item 3	Item 4	Total par item
Item 1	0,24	-0,12	0	0,12	0,24
Item 2	-0,12	0,2	0	0,08	0,16
Item 3	0	0	0	0	0
Item 4	0,12	0,08	0	0,2	0,4
Items 1 à 4					0,8

1.3 MOYENNE ET VARIANCE D'UN SCORE COMPOSITE

Nous pouvons donc conclure de l'observation de ces deux exemples que la moyenne d'un score composite est la somme des moyennes de ses composantes, tel qu'exprimé dans l'équation suivante :

$$\bar{X}_c = \sum \bar{X}_i \tag{4.2}$$

Quant à la variance du score composite, nous pouvons démontrer qu'elle est égale à la somme des variances et des covariances des items qui composent le test. Comme nous l'avons vu également dans les tableaux 2 et 3, la relation entre la variance d'un score total et la variance de ses composantes n'est pas aussi simple. En plus de la variance des items individuels, la *covariance* entre les items joue un rôle important dans l'estimation de la variance du score total.

Il est facile de constater que la variance totale des scores est constituée de la somme des variances et covariances entre les items. En effet, si l'on additionne tous les éléments de la matrice des variances covariances du tableau 3, on retrouve essentiellement la même valeur de variance que celle calculée pour le test constitué des quatre items (0, 82 ou 0, 8 selon le degré de précision des calculs).

Si l'on compare à nouveau les tests des tableaux 2 et 3, on s'apercevra que la variance totale du test composé des trois items 1, 2 et 4 est bien supérieure lorsqu'il s'agit du premier test que lorsqu'il s'agit du second test. Elle est de 2,20 dans le premier test et de 0,82 dans le second test. La variance est dans ce cas-ci presque trois fois supérieure dans le test homogène que dans le test hétérogène. L'impact de l'item 4 est également intéressant. Dans le premier test, l'item 4 a contribué à augmenter la variance totale de 0,98 à 2,20 (deux fois plus). Dans le second test, l'item 4 a contribué à faire passer la variance totale de 0,20 à 0,82 (quatre fois plus).

La variance totale d'un test dépend donc non seulement de la variance de ses items individuels, mais aussi de leur homogénéité. Plus la covariance entre les items est élevée, plus la variance totale au test sera grande. En fait, il est possible de démontrer que la variance totale à un test est le résultat de la somme des variances des items et des covariances entre items. L'encadré 1 fait la preuve de cet énoncé.

Encadré 1

Soit un test C composé de deux scores X_1 et X_2. Transformons ces scores en scores centrés à la moyenne, c, x_1, x_2 afin de simplifier les calculs de la variance et de la covariance.

$$c = x_1 + x_2 \tag{4.3}$$

L'expression de la variance du score composite *c* est donnée par les équations suivantes, la seconde substituant *c* par ses valeurs.

$$\sigma_c^2 = \frac{\sum c^2}{n} \quad (4.4)$$

$$\sigma_c^2 = \frac{\sum (x_1 + x_2)^2}{n} \quad (4.5)$$

Si l'on développe la dernière expression, l'on obtient :

$$\sigma_c^2 = \frac{\sum x_1^2 + 2x_1x_2 + x_2^2}{n} \quad (4.6)$$

En répartissant la sommation sur chaque membre de l'addition, on peut réécrire l'équation (4.6), de la façon suivante :

$$\sigma_c^2 = \frac{\sum x_1^2}{n} + 2\frac{\sum x_1x_2}{n} + \frac{\sum x_2^2}{n} \quad (4.7)$$

Le premier et le dernier terme de l'addition ne sont autres que l'expression de la variance des items (exprimée en scores centrés). Le terme du centre est l'expression de la covariance entre les items. On peut donc reformuler l'équation (5) en termes de variances et de covariances :

$$\sigma_c^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} \quad (4.8)$$

Dans le cas de tests possédant un nombre j d'items, il est possible de généraliser la démonstration précédente de manière à prouver que :

$$\sigma_c^2 = \sum \sigma_i^2 + 2\sum \sigma_{ij} \quad (4.9)$$

où $\sum \sigma_i^2$ représente la somme des variances des items et $\sum \sigma_{ij}$, la somme des covariances des items pris deux à deux.

La figure 1 représente la matrice de variances-covariances entre items. Il s'agit d'une matrice carrée symétrique. Les variances de chaque item figurent en diagonale et les covariances de part et d'autre de la diagonale principale. Puisque la covariance *ij* est la même que la covariance *ji*, les mêmes valeurs se répètent symétriquement par rapport à la diagonale de la matrice.

Cette figure permet de nous rendre compte que la covariance entre les items joue un rôle proportionnellement beaucoup plus important que la variance des items individuels dans la variance totale des résultats à un test. Un test comptant *j* items sera le résultat de la somme de *j* variances d'items et de *j(j-1)* covariances. Si, ainsi l'on ajoute 10 items à un test en comportant déjà 10, sa variance totale sera augmentée de la variance individuelle de 10 items mais aussi de 380 covariances (20x19) comparativement à 90 (10x9) au départ. C'est donc dire que ces 10 nouveaux items contribueront à accroître la variance totale des résultats au test dans la mesure où ils covarient de manière importante avec les items déjà présents. Pour cela, les items ajoutés doivent constituer un ensemble homogène avec les items de départ.

$$\sigma_c^2 = \Sigma \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{bmatrix}$$

Figure 1 – Matrice des variances-covariances et variance totale

1.4 Implications pour la construction d'un test

Des observations précédentes, il ressort trois conséquences principales pour la construction d'un test :

- augmenter le nombre d'items accroît la variance totale d'un test dans la mesure où les items supplémentaires sont homogènes avec les items déjà présents dans le test ;
- les items ayant un contenu similaire sont plus susceptibles d'avoir une covariance élevée et ainsi de contribuer davantage à la variance totale des résultats au test ;
- pour contribuer de façon significative à la variance totale du test, l'item doit de préférence être de difficulté moyenne : un item trop facile ou trop difficile n'a qu'une faible variance et une faible covariance.

Lorsque l'objectif est de différencier les personnes, ces observations nous mettent en garde contre la tentation d'inclure trop d'items différents et sans rapport entre eux dans le score total d'un individu. Par exemple, l'enseignant qui souhaite mieux différencier ses élèves en ajoutant de nouveaux items dans son examen, aura avantage à prendre des items supplémentaires évaluant les mêmes objectifs que ceux déjà évalués par le test initial. S'il choisit au contraire de faire porter les items supplémentaires sur de nouveaux objectifs sans lien avec ceux déjà évalués, il risque de gagner bien peu en différenciation des scores des élèves. Il y aurait alors avantage à calculer deux scores totaux séparés et à établir un profil de scores.

Une bonne variance des résultats est une condition nécessaire quoique non suffisante pour obtenir des résultats fiables et valides. Sans anticiper sur les prochaines sections, il est important de faire ressortir que dans le contexte d'une évaluation normative, la variance joue un rôle important. Comment sélectionner les meilleurs élèves pour un cours d'art plastique si les résultats sur lesquels on doit se baser sont semblables les uns aux autres et ne permettent pas de les différencier ?

Si l'objectif n'est pas de différencier les sujets, alors il n'est pas aussi essentiel d'obtenir une variance élevée des résultats. Il existe des situations où celle-ci n'est pas une condition importante, comme dans le cadre de la pédagogie de la maîtrise où l'on prévoit que la presque totalité des élèves atteidront les objectifs d'apprentissage. Dans

de telles circonstances, on ne s'attend pas à ce que l'examen puisse établir de différences entre les individus. Le but consiste plutôt à faire disparaître ces différences par un enseignement approprié.

Enfin, au-delà des strictes considérations de variance, il est important de se rappeler qu'il peut être beaucoup plus facile d'interpréter des résultats homogènes que des résultats hétérogènes. Il y a peu d'intérêt à différencier les personnes si l'on ignore en quoi exactement ils sont différents. Si un score total est un ensemble d'items hétérogènes, alors il devient presqu'impossible d'identifier les causes véritables qui font de chaque examiné un individu différent des autres.

2. La théorie classique des scores

C'est Spearman (1907) qui a jeté les fondements de la théorie classique des scores. La théorie, dans sa forme actuelle, est due principalement aux travaux de Gulliksen (1950), Magnusson (1967) et de Lord et Novick (1968). C'est sous sa formulation la plus récente que nous en discuterons dans les sections suivantes.

2.1 Postulats du modèle

La théorie classique permet de répondre simplement à plusieurs des questions précédentes. Elle est sans doute le modèle le plus simple de ceux que nous verrons. Ce modèle a l'avantage de pouvoir être utilisé dans une grande variété de situations parce que ses postulats de départ sont faibles — au sens d'aisés à satisfaire — et peu nombreux, ce qui n'est pas le cas des modèles de la réponse aux items.

La théorie classique des tests est aussi appelée « *théorie classique des scores* » puisque son objet d'intérêt est le score total obtenu par une personne à un test. Les postulats principaux de la théorie classique sont :

Postulat 1 :

Le score observé d'un individu résulte de la somme entre le score vrai de l'individu (V : une constante) et l'erreur de mesure associée à ce score (E : une variable aléatoire) :

$$X = V + E \tag{4.10}$$

Il résulte de cette équation que le score observé X est également une variable aléatoire. Par exemple, si le score vrai d'un élève est 84%, il est possible que celui-ci obtienne 87% ou 76% à un examen. Toutefois, la probabilité d'obtenir un score différent de 84% décroît au fur et à mesure que l'on s'éloigne du score vrai. En fait, l'erreur de mesure se distribue normalement, ce qui fait que le score observé lui-même se distribue normalement autour du score vrai. C'est ce qui est illustré par la figure 2.

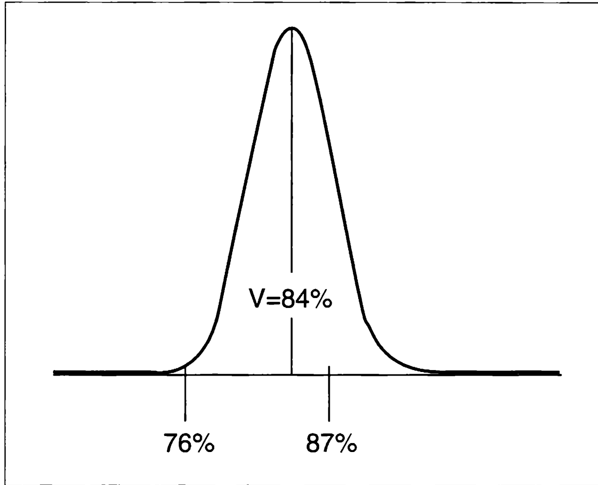


Figure 2 – Distribution théorique des scores observés autour du score vrai.

Postulat 2 :

Il est conséquent avec le premier. Il stipule que la valeur attendue pour le score observé est le score vrai :

$$\varepsilon(X) = V \tag{4.11}$$

Ce postulat signifie tout simplement que le score vrai d'un individu est l'espérance mathématique des scores observés. En d'autres mots, la précision d'un score observé s'accroît avec le nombre d'observations. En effet, si l'on devait administrer plusieurs fois le même test au même sujet, la moyenne des résultats nous fournirait, à la limite, son score vrai. Le score vrai peut ainsi être considéré comme la moyenne de la distribution théorique des scores observés du sujet, en supposant qu'il soit possible d'administrer de manière indépendante le même test à plusieurs reprises au même sujet. La dispersion des scores observés X autour du score vrai V constitue l'erreur de mesure pour l'ensemble de ces passations.

Postulat 3 :

Il n'y a pas de corrélation entre l'erreur de mesure et le score vrai de l'individu :

$$\rho_{EV} = 0 \tag{4.12}$$

Ceci signifie, par exemple, que l'erreur aléatoire de mesure ne sera pas plus grande si l'individu possède un score vrai élevé ou plus faible s'il possède un score vrai faible. Une telle situation se produirait, par exemple, si un enseignant corrigeait plus attentivement les copies des élèves faibles que les copies des élèves forts et que, par conséquent, les erreurs de correction seraient plus importantes chez les élèves forts (corrigées plus rapidement) que chez les élèves faibles (corrigées plus attentivement).

Postulat 4 :

Ce postulat stipule que les erreurs à deux tests différents (E_1 et E_2) ne sont pas corrélées entre elles :

$$\rho_{E_1 E_2} = 0 \qquad (4.13)$$

Ceci peut se produire lorsque, par exemple, un sujet fatigué obtient des notes plus faibles à différents tests administrés en fin de journée. Dans ce cas, les erreurs sont liées entre elles puisqu'elles résultent d'un même facteur sous-jacent.

Postulat 5 :

Il n'y a pas de corrélation entre l'erreur de mesure à un test et le score vrai à un autre test :

$$\rho_{E1V2} = 0 \qquad (4.14)$$

Supposons qu'un questionnaire à choix de réponse mesure la créativité et que, plus le score de créativité est élevé, plus l'élève est porté à répondre au hasard lorsqu'il ignore la réponse. Dans cette situation, il y aurait une corrélation entre le score vrai au test et l'erreur aléatoire de mesure qui ne serait pas la même pour les individus créatifs que pour les individus non créatifs. En fait, on peut affirmer que le postulat 5 ne tient pas dès que le test mesure une caractéristique de l'individu qui exerce une influence directe ou indirecte sur sa façon de répondre au test, telle que la tendance à deviner, à tricher, à omettre certaines catégories de réponses, etc.

Postulat 6 :

Deux tests X et X' sont parallèles si et seulement si leurs scores vrais et leurs erreurs de mesure sont égales :

$$V = V' \qquad (4.15)$$

$$\sigma_E = \sigma_{E'}$$

À cause du postulat 1 qui stipule que le score observé est la somme d'un score vrai et d'un score d'erreur aléatoire, il découle que deux tests parallèles auront sensiblement la même moyenne et la même variance des scores observés.

Postulat 7 :

Il définit ce qu'est un test τ-équivalent (prononcer « tau-équivalent »). Deux tests sont considérés comme τ-équivalents lorsque leurs scores vrais diffèrent par une constante additive *k*.

$$V_1 = V_2 + k \qquad (4.16)$$

Ainsi, si trois sujets obtiennent 10, 23 et 19 à un test et qu'ils obtiennent 17, 30 et 26 à un autre test, ces deux tests sont τ-équivalents, la constante *k* valant 7. Il découle de cette dernière définition que les tests parallèles rencontrent les exigences des tests τ-équivalents, alors que la réciproque n'est pas vraie.

2.2 Implications de la théorie classique des scores

L'ensemble des sept postulats de la théorie classique se résume facilement : les erreurs aléatoires de mesure doivent être indépendantes en toutes circonstances. Ceci signifie que les conditions de testing doivent être telles qu'il n'y a pas de corrélation entre le score vrai d'un sujet et l'erreur de mesure, ni entre l'erreur de mesure à un test et l'erreur de mesure à un autre test. Ce sont là des conditions minimales sans lesquelles les scores observés deviennent difficilement interprétables. Par exemple, pour démontrer que deux items à la fin d'un test mesurent bien la même caractéristique, il faut écarter l'hypothèse que la corrélation entre ces deux items puisse être le résultat d'erreurs de mesure dues à la fatigue, à l'ennui ou à un manque de motivation.

La théorie classique tient compte d'une erreur strictement aléatoire. Si les postulats de base sont respectés, c'est-à-dire si les différentes sources d'erreur sont indépendantes les unes des autres, alors celles-ci pourront s'annuler de sorte que sur un grand nombre de mesures répétées, l'espérance mathématique des scores observés soit le score vrai de l'individu. Si ces erreurs ne sont pas indépendantes, alors leurs effets risquent d'être non nuls et l'équation de départ (postulat #1) est inadéquate pour représenter la situation que l'on cherche à décrire.

D'autres sources d'erreur peuvent invalider nos résultats. Il s'agit de sources d'erreur dont l'effet est constant et dont la résultante est non nulle : les erreurs systématiques. Ces sources d'erreur ne sont pas prises en ligne de compte par la théorie classique et doivent faire l'objet d'une étude particulière : la validation des résultats. Par exemple, il y a erreur systématique lorsqu'un test est trop facile ou trop difficile. Deux sujets, dont les scores vrais en mathématiques sont différents, peuvent obtenir le même score vrai de 10/10 lorsque l'examen est trop facile. De manière identique, des sujets handicapés auditifs ou handicapés visuels verront leurs scores vrais en orientation spatiale systématiquement sous-estimés par des épreuves sensorielles auditives ou visuelles. Comme le handicap est permanent, celui-ci fait partie du score vrai de l'élève à l'épreuve en question. C'est pourquoi on pourrait réécrire l'équation (4.10) de la manière suivante :

$$X = V + e_s + e_a \qquad (4.17)$$

Dans cette dernière expression, le score observé du sujet est la somme d'un score vrai, d'une erreur systématique e_s et d'une erreur aléatoire e_a. Par exemple, un enseignant qui ferait porter une grande part de son examen sur une partie sans importance de la matière ou encore sur des objectifs dont les élèves n'ont pas été informés, mesurerait davantage ce qui n'est pas pertinent. Dans un tel contexte, l'erreur aléatoire de mesure serait dérisoire par rapport à l'erreur systématique introduite. Bref, ce test fournirait des résultats précis (fiables), mais sans grande validité.

En fait, il est possible de représenter les notions de validité et de fiabilité en fonction de la proportion de la variance des scores observés imputable à de la variance pertinente (σ_v^2), à de la variance non pertinente ($\sigma_{e_s}^2$) ou à de la variance d'erreur ($\sigma_{e_a}^2$). La figure suivante explique le rapport entre ces différentes variances.

De la figure 3, il ressort que le but du constructeur de test est de maximiser la part du score vraie qui est pertinente à ce qu'il souhaite mesurer, tout en minimisant l'erreur aléatoire de mesure. Pour ce faire, il faut que la variance des scores vrais occupe une grande proportion de la variance des scores observés et que la variance d'erreur systématique soit minimale. Des tests 1 à 3 de la figure 3, le test 3 est le plus fiable et le plus valide : c'est celui pour lequel la proportion de la variance des scores observés qui soit de la variance imputable aux scores vrais est la plus élevée. Par contre, le test 3 n'est pas plus fiable que le test 2, car tous deux comportent la même proportion de la variance des scores observés qui est de la variance d'erreur aléatoire. Enfin, le test 1 est le moins fiable des trois tests. Une grande proportion de la variance des scores observés provient de la variance d'erreur aléatoire.

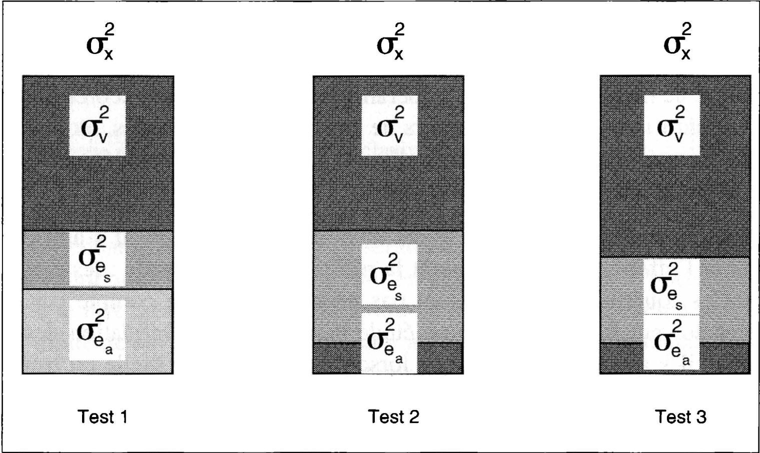


Figure 3 – Répartition des variances d'erreur dans trois tests différents

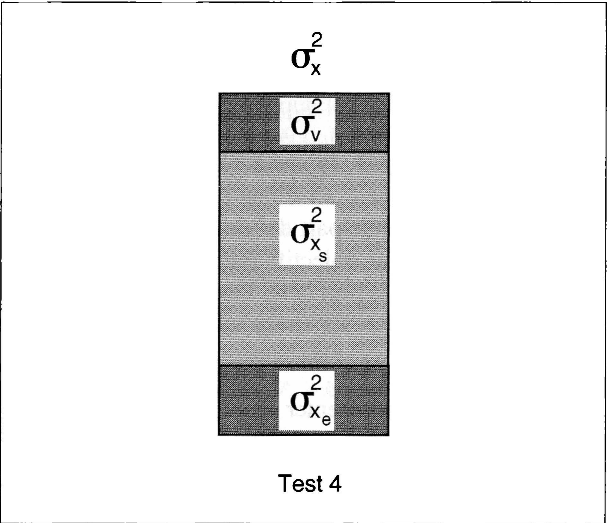


Figure 4 – Répartition de la variance d'erreur dans un test peu valide mais fidèle

Il ressort aussi de la figure 4 qu'un test peut-être fiable même s'il ne mesure pas ce que l'on souhaite. Dans ce cas-ci, la proportion de la variance observée constituée d'erreur systématique est très importante. Ceci assure une grande précision et une grande stabilité des scores, mais pas sans rapport avec la caractéristique que nous souhaitons mesurer (p.e. : un test trop facile). Il est donc possible qu'un instrument fiable ne mesure pas ce que nous souhaitons mesurer. De bons indicateurs de fiabilité, même s'ils permettent d'envisager que les résultats au test puissent être valides, ne sont pas une condition suffisante à eux seuls.

Prenons pour exemple une mesure bien connue de la compétition sportive : le tir à l'arc. Lorsque, comme l'indique la situation *A* de la figure 5, un tir groupé rate systématiquement la cible, on peut parler de tir fiable, mais pas de tir valide. Il suffira au tireur de corriger le biais de son tir pour le rendre valide et ainsi d'atteindre le *1000* ou *l'oeil de boeuf.* Par contre, un instrument non fiable ne saurait être valide, à cause du manque de précision de la mesure. C'est le cas d'un tir dispersé sur toute la surface de la cible (situation *B*). Le problème de ce tireur est beaucoup plus délicat que le premier. Comme son tir est aléatoire et non systématique, il y a peu de choses que nous puissions faire pour l'aider à corriger son tir. Un premier bon pas dans cette direction serait de s'assurer que le tireur possède au moins un tir groupé. Enfin, la situation *C* représente la situation souhaitée : un tir groupé qui touche le mille.

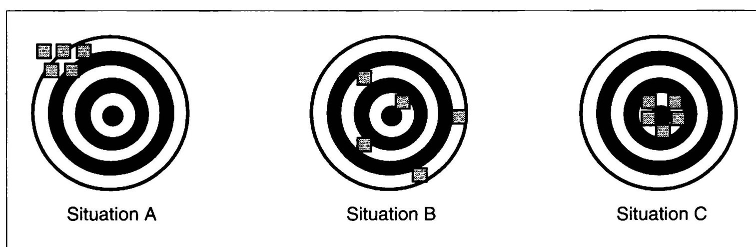


Figure 5 – Trois tirs à la cible différents en termes de validité et de fiabilité

Ce dernier exemple illustre pourquoi la fiabilité est un indicateur si important de la qualité des résultats. Sans fiabilité ou, si l'on préfère, sans mesure précise, toute discussion sur la validité devient futile et tout espoir de rectifier notre tir est vain tant que les résultats manqueront de précision.

Si la fiabilité est si importante, que peut-on faire pour l'augmenter ? Disons pour l'instant que le nombre d'items joue un rôle important dans la précision de la mesure. Plus il y a d'items entrant dans le calcul du score observé, plus celui-ci risque d'être précis à condition que ces items mesurent bien la même chose. Ceci découle du postulat 2 qui indique que la moyenne des scores observés d'un individu tend vers son score vrai. Plus il y aura d'items, plus l'erreur type d'estimation de cette moyenne sera faible et, par conséquent, plus l'erreur de mesure sera réduite. Il y a lieu de souligner que la variance des scores vrais augmente plus vite que la variance d'erreur lorsqu'on ajoute de nouveaux items à un test.

2.3 DÉFINITIONS DE LA FIABILITÉ

Le praticien a besoin d'avoir une idée de l'écart qui existe entre la note obtenue et la note vraie. La fiabilité nous renseigne sur le degré de relation entre les deux notes. Il est possible de formuler plusieurs définitions de la fiabilité à partir des sept postulats de la théorie classique. Certaines de ces définitions n'ont qu'un intérêt théorique. D'autres ont un intérêt pratique car elles nous permettent d'estimer la fiabilité d'un test s'il y a de bonnes raisons de croire que le modèle de la théorie classique s'applique à nos résultats. Nous verrons les trois définitions générales suivantes :

L'INDICE DE FIABILITÉ : c'est la corrélation entre les scores observés et les scores vrais. Lorsque cette corrélation est égale à 1 (fiabilité parfaite), scores vrais et observés sont égaux et il n'y a pas d'erreur de mesure. Lorsque cette corrélation est égale à 0, alors chaque score vrai peut correspondre à n'importe quel score observé et l'erreur de mesure devient égale à l'écart type des scores observés. L'écart type des scores observés est en effet la plus grande erreur de mesure possible. Voici la représentation symbolique de cette définition de l'indice de fiabilité :

$$\rho_{xy} = \sum_{i=1}^{N} \frac{xv}{N\sigma_x\sigma_v} \qquad (4.18)$$

L'équation (4.17) représente le calcul de la corrélation pour des scores centrés (écarts à la moyenne). Elle n'a aucun intérêt pratique puisque nous ne connaissons pas la valeur du score vrai *v*. Sur le plan conceptuel, elle nous permet de comprendre cependant que meilleure est la fiabilité, meilleure sera la prédiction du score vrai à partir du score observé.

LE COEFFICIENT DE FIABILITÉ *(DÉFINITION THÉORIQUE)* : c'est la proportion de la variance des scores observés qui est imputable aux scores vrais. Elle signifie que plus le test est précis, plus la variance des scores observés est due à la variance des scores vrais et non à des fluctuations du hasard. Concrètement, le coefficient de fiabilité $r_{xx'} = 0{,}81$ signifie que 81% de la variance des scores observés est attribuable à la variance des scores vrais. L'équation (4.19) illustre cette relation.

$$\rho_{xx'} = \frac{\sigma_v^2}{\sigma_x^2} \qquad (4.19)$$

LE COEFFICIENT DE FIABILITÉ *(DÉFINITION OPÉRATIONNELLE)* : c'est la corrélation entre scores observés à deux formes parallèles. Puisque par définition, le score vrai d'un même sujet à deux tests parallèles est le même, la corrélation entre les scores observés à deux tests parallèles nous fournit par le fait même la proportion de la variance des scores observés qui est la variance du score vraie. En effet, la corrélation des scores vrais avec eux-mêmes nous fournit la variance des scores vrais et comme nous connaissons déjà la variance des scores observés, il est facile d'estimer la fiabilité du test en calculant la proportion de la variance des scores observés qui provient de la variance du score vrai. Bref, les tests parallèles nous fournissent une méthode pour opérationnaliser l'estimation de la variance des scores vrais lorsque les postulats de base de la théorie classique sont raisonnablement satisfaits.

$$\rho_{x_1x_2} = \sum \frac{x_1x_2}{N\sigma_{x_1}\sigma_{x_2}} \tag{4.20}$$

Le lecteur intéressé pourra prendre connaissance de la démonstration dans l'encadré 2 qui illustre comment l'équation (4.20) est bien une opérationnalisation de l'équation (4.19).

Encadré 2

Au départ, nous pouvons poser que la corrélation entre les scores observés à deux tests parallèles est représentée par l'équation suivante, représentant la corrélation entre les scores centrés x_1 et x_2 aux deux tests parallèles.

$$\rho_{x_1x_2} = \sum \frac{x_1x_2}{N\sigma_{x_1}\sigma_{x_2}} \tag{4.21}$$

Or, nous savons à partir du postulat #1 de la théorie classique que x_1 et x_2 peuvent être exprimés en termes de scores vrais v_1 et v_2.

$$x_1 = v_1 + e_1 \tag{4.22}$$

$$\text{et } x_2 = v_2 + e_2 \tag{4.23}$$

Nous pouvons développer l'expression de la corrélation entre deux tests parallèles en remplaçant les scores observés x_1 et x_2 par leur valeur exprimée en scores vrais. Ceci nous donne l'expression suivante :

$$\rho_{x_1x_2} = \sum \frac{(v_1 + e_1)\ (v_2 + e_2)}{N\sigma_{x_1}\sigma_{x_2}} \tag{4.24}$$

Une fois la multiplication développée au numérateur, nous obtenons l'expression :

$$\rho_{x_1x_2} = \sum \frac{v_1v_2 + v_1e_2 + v_2e_1 + e_1e_2}{N\sigma_{x_1}\sigma_{x_2}} \tag{4.25}$$

Enfin, suite à la distribution de la sommation, l'expression de la corrélation entre scores centrés à deux tests parallèles prend la forme suivante :

$$\rho_{x_1x_2} = \sum \frac{v_1v_2}{N\sigma_{x_1}\sigma_{x_2}} + \sum \frac{v_1e_2}{N\sigma_{x_1}\sigma_{x_2}} + \sum \frac{v_2e_1}{N\sigma_{x_1}\sigma_{x_2}} + \sum \frac{e_1e_2}{N\sigma_{x_1}\sigma_{x_2}} \tag{4.26}$$

Cette dernière expression peut maintenant être grandement simplifiée. Les postulats 4 et 5 stipulent en effet les relations suivantes :

$$\rho_{e_1e_2} = 0 \tag{4.27}$$

$$\rho_{v_1e_2} = \rho_{v_2e_1} = 0 \tag{4.28}$$

Ceci permet donc d'égaler à 0 les trois derniers termes de l'addition de l'équation 6, puisqu'il s'agit en fait de la corrélation entre erreurs de mesure (4.26) et de la corrélation entre score vrai et erreur de mesure (4.27). L'équation devient donc la suivante :

$$\rho_{x_1x_2} = \sum \frac{v_1 v_2}{N\sigma_{x_1}\sigma_{x_2}} \quad (4.29)$$

Cette équation peut être réécrite en tenant compte du postulat 6 de la théorie classique qui définit les propriétés des tests parallèles de la manière suivante :

$$v_1 = v_2 \quad (4.30)$$

$$\sigma_1 = \sigma_2 \quad (4.31)$$

En substituant ces relations dans l'équation (4.29), il est possible d'écrire :

$$\rho_{x_1x_2} = \sum \frac{V^2}{N\sigma_{x_1}\sigma_{x_2}} = \frac{\sigma_v^2}{\sigma_x^2} \quad (4.32)$$

car,

$$\sum \frac{v^2}{N} = \sigma_v^2 \quad (4.33)$$

En fait, ce dernier développement est possible car la variance des scores observés à chaque test parallèle est égale par définition et que les scores vrais à chaque test parallèle sont les mêmes (4.29 et 4.30). Dans ce dernier cas, la covariance entre scores vrais à deux tests parallèles est égale à la variance des scores vrais.

Cette démonstration illustre que la corrélation entre deux formes parallèles d'un test nous permet d'en estimer la fiabilité pour autant que les postulats de la théorie classique soient adéquats pour décrire nos résultats.

3. Estimation de la fiabilité

3.1 Méthode des formes parallèles

La définition opérationnelle de la fiabilité nous indique dans quelles conditions la théorie classique nous permet d'estimer la précision des scores. À condition de disposer de deux formes parallèles, il est possible de calculer la proportion de la variance des scores observés qui est due aux scores vrais et ainsi d'estimer la fiabilité lorsque les postulats de la théorie classique sont valables. Dans la pratique, ces deux formes parallèles peuvent se rencontrer dans les situations que voici :

La stabilité. Si l'on administre le même test à deux reprises, la corrélation entre les scores observés au test-retest nous donne une indication de la stabilité des résultats dans le temps. Le test administré au temps A est considéré comme parallèle au même test administré au temps B. Si les résultats au test-retest ne sont pas stables, alors la corrélation entre les deux sera faible et l'effet du passage du temps s'ajoutera à l'erreur de mesure. Il faut noter qu'une telle procédure suppose que le retest est sans effet particulier sur les sujets, c'est-à-dire qu'il n'y a pas eu d'effet d'apprentissage ou de contamination des résultats. Si, par exemple, les sujets les plus forts lors de la première administration sont aussi ceux qui, au moment du retest, se rappellent mieux des questions posées la première fois, il risque d'y avoir corrélation entre le score vrai de

l'élève au premier test et l'erreur aléatoire de mesure au second, ce qui enfreint le postulat 5 du modèle de la théorie classique.

L'ÉQUIVALENCE. Si l'on administre deux versions d'un même test, la corrélation entre les scores de chaque test nous renseigne sur le degré d'équivalence entre les tests. Ceci suppose bien entendu que les deux formes ont été administrées en même temps ou à l'intérieur d'une période de temps très courte, sinon la stabilité et l'équivalence des deux tests seraient mesurées simultanément. Ce type de fiabilité requiert que deux tests soient créés. Ce n'est pas toujours nécessaire cependant. On peut décider de considérer comme équivalentes les deux moitiés d'un test (méthode de bissection). Le calcul de la corrélation nous fournit alors une estimation de l'équivalence des résultats pour chaque moitié du test. À la limite, on peut étendre ce concept jusqu'aux items et déterminer à quel point tous les items entrant dans le calcul d'un score total sont homogènes, c'est-à-dire équivalents ou encore, en langage de la théorie classique, parallèles. L'ennui avec le calcul de la fiabilité par la méthode de bissection, c'est que l'estimation fournie se fonde sur une partie du test, alors que c'est la fiabilité du test entier que nous recherchons. Des corrections visant à tenir compte de cette situation sont disponibles (Spearman-Brown, Guttman, Rulon) et nous les verrons dans la section 3.2.

LA STABILITÉ-ÉQUIVALENCE. Dans le calcul de la stabilité, on cherche à déterminer l'effet du passage du temps sur la fiabilité du score. Dans le calcul de l'équivalence, c'est l'effet de l'échantillonnage des items sur le score total de l'individu que l'on cherche à mesurer. Lorsque l'on cherche à tenir compte de ces deux sources de fluctuation du score total, nous procédons au calcul d'un coefficient de stabilité-équivalence. Cette valeur de fiabilité nous est fournie par la corrélation entre les deux formes d'un test à des moments différents. En fait, le calcul de la stabilité-équivalence devient nécessaire lorsque l'on ne peut utiliser le même test dans le calcul de la stabilité. Comme deux sources de fluctuation aléatoire seront présentes dans le calcul de cette corrélation, le coefficient de stabilité-équivalence est généralement la plus faible estimation de la fiabilité parmi les trois que nous venons d'envisager.

Le tableau 4 décrit le plan d'observation des trois méthodes précédentes de calcul de la fiabilité. Il s'agit en fait de deux tests parallèles (test 1 et test 2) administrés à deux moments différents (temps A et temps B). Il est donc possible d'estimer la fiabilité de trois manières différentes :

- la stabilité : par la corrélation entre les résultats de chaque test (test 1 ou test 2) au temps A avec les résultats au même test au temps B ;
- l'équivalence : par la corrélation entre les résultats des deux tests parallèles administrés au même moment (soit au temps A, soit au temps B) ;
- la stabilité-équivalence : par la corrélation entre les résultats aux deux tests parallèles administrés à des moments différents (test 1 au moment A avec test 2 au moment B ; test 2 au moment A avec test 1 au moment B).

Le tableau 4 présente un exemple de calcul pour chaque méthode d'estimation de la fiabilité. Il présente la moyenne et l'écart type de chaque test aux moments A et B, ainsi que les corrélations nécessaires à l'estimation de la fiabilité des résultats.

Tableau 4 – Fiabilité de stabilité et équivalence : exemples de calculs

Temps A			Temps B		
Sujets	Test 1	Test 2	Sujets	Test 1	Test2
1	9	8	1	13	11
2	7	8	2	8	9
3	7	7	3	8	8
4	8	6	4	8	6
5	5	6	5	7	6
6	4	4	6	5	5
7	5	5	7	4	4
Moyenne	6,43	6,29	Moyenne	7,57	7
Variance	2,82	1,92	Variance	7,1	5,14
Corr.	0,81		Corr.	0,92	

Stabilité/équivalence

		Temps A	TempsA
		Test 1	Test 2
Temps B	Test 1	0,87	0,81
Temps B	Test 2	0,79	0,91

Équivalence

	Temps A	
	Test 1	Test 2
Test 1	1	0,81
Test 2	0,81	1

	Temps B	
	Test 1	Test 2
Test 1	1	0,92
Test 2	0,92	1

Premièrement, on constate que les moyennes des deux tests parallèles ne sont pas exactement les mêmes, de même que les écarts types. Est-ce une raison suffisante pour remettre en question le modèle de la théorie classique ? D'abord, il ne faut pas être étonné de différences entre les moyennes et les écarts types à cause des fluctuations d'échantillonnage. Ensuite, il faut se demander si ces fluctuations sont telles qu'elles remettent en question l'hypothèse selon laquelle les tests sont parallèles. Rappelons que l'équivalence entre les tests signifie que les moyennes et les écarts types des résultats à ces tests sont les mêmes dans la *population*, en dépit des différences observées au niveau de l'*échantillon*.

Les résultats indiquent que les tests 1 et 2 sont relativement équivalents, que l'estimation de l'équivalence entre les deux tests ait été faite au temps A ou au temps B. En effet, la corrélation entre les deux tests est de 0, 81 au temps A et de 0, 92 au temps B. Il y a donc entre 65 et 85% de la variance qui est commune aux deux tests, c'est-à-dire de la variance imputable à la variance des scores vrais si le modèle de la théorie classique est approprié. Ces deux pourcentages de variance nous sont donnés par le calcul du *coefficient de détermination* qui consiste à élever au carré la corrélation entre les deux tests. C'est ainsi que $r^2 = (0{,}81)^2 \cong 65\%$ au temps A et que $r^2 = (0{,}92)^2 \cong 85\%$ au temps B.

Les résultats à chacun des tests sont également très stables. La stabilité des résultats du test 1 est estimée à 0,87 et celle du test 2 à 0,91. Il y a donc respectivement 76% et 83% de variance commune entre la première et la seconde administration du test 1 et du test 2. Ceci indique encore une fois qu'une grande proportion de la variance des scores observés est imputable aux scores vrais, et que les résultats demeurent relativement précis lors de mesures répétées dans le temps sur les mêmes personnes.

La plus faible valeur de fiabilité est celle de la stabilité-équivalence. En effet, cette estimation cumule les erreurs aléatoires de mesure imputables aux différences d'échantillonnage des items entre les deux tests parallèles, de même que les erreurs aléatoires de mesure imputables à l'effet du temps. La valeur estimée ne saurait donc être plus grande que la plus petite des valeurs de fiabilité précédemment calculée, qu'il s'agisse d'équivalence ou de stabilité. La stabilité-équivalence du test 1 est de 0,81 et celle du test 2 est de 0,79. Dans le cas du test 1, la valeur de stabilité-équivalence est égale au coefficient d'équivalence avec le test 2 mesurée au temps A (0,81=0,81), mais inférieure au coefficient de stabilité (0,81<0,87) et au coefficient d'équivalence mesuré au temps B (0,81<0,92). Dans le cas du test 2, la valeur de stabilité-équivalence est inférieure à la fois à l'équivalence du test 2 avec le test 1 (temps A : 0,79<0,81 ; temps B : 0,79<0,92) et à la stabilité test-retest (0,79<0,91). Le coefficient de stabilité-équivalence est donc, parmi les trois méthodes précédentes, celle qui fournit l'estimation de la fiabilité la plus basse.

3.2 MÉTHODE DE BISSECTION

Tous les calculs pratiques de la fiabilité que nous avons pris en considération jusqu'à présent possèdent un point en commun : ils requièrent la construction de deux tests ou encore l'administration du même test à deux reprises. Aucune de ces trois méthodes ne permet d'obtenir une estimation de la fiabilité avec un seul test qui aurait fait l'objet d'une seule administration.

Pourtant, dans le cas précis de l'équivalence entre deux tests, il est possible de considérer des parties quelconques du test entier comme des formes parallèles du même test. Le problème de la double administration et de la construction de deux tests est ainsi contourné. Cette méthode est connue sous le nom de la méthode de bissection (en anglais « *split-half* »).

Mais que mesure au juste la corrélation entre les scores totaux à deux moitiés d'un test ? Il s'agit en fait d'une mesure de la cohérence interne du test entier, ou si l'on préfère, de l'homogénéité de ses parties. Si la corrélation obtenue entre les deux moitiés est élevée, elle indique que les deux parties mesurent sensiblement la même chose. Si, par contre, elle est faible, elle indique qu'elles mesurent des choses différentes. En fait, la cohérence interne du test nous permet de nous prononcer sur l'additivité des scores entre les différentes parties d'un test. Si ces parties mesurent des caractéristiques différentes, sans liens ou encore en relations opposées, leur addition contribuera bien peu à améliorer la fiabilité du score total au test. La variance du score total ainsi obtenue sera considérablement réduite, comme nous l'avons déjà démontré dans les exemples des tableaux 2 et 3.

Tableau 5 – Méthode de bissection

Sujets#	Items#										Items pairs	Items impairs	Items 1 à 5	Items 6 à 10
	1	2	3	4	5	6	7	8	9	10				
1	2	2	3	1	3	4	1	2	4	1	10	13	11	12
2	4	3	3	1	3	5	2	4	5	2	15	17	14	18
3	5	3	5	3	4	5	3	5	5	2	18	22	20	20
4	2	1	3	2	3	4	1	3	5	1	11	14	11	14
5	3	2	5	4	5	5	2	4	5	1	16	20	19	17
7	1	1	3	2	3	3	0	2	4	2	10	11	10	11
8	1	1	3	1	4	3	0	1	3	0	6	11	10	7
9	0	1	2	0	3	3	0	1	3	0	5	8	6	7
10	5	3	5	3	3	5	3	5	5	3	19	21	19	21
Moyenne	2,6	2	3,7	2,1	3,6	4,2	1,4	3,1	4,4	1,3	12,7	15,7	14	14,4
Variance	2,6	0,8	1,2	1,7	0,6	0,8	1,2	2,1	0,6	0,8	22,4	22	23,6	22,8
Corrélation entre les deux moitiés « pairs/impairs » : 0,973														
Corrélation entre les deux moitiés « premies/derniers » : 0,887														

Les méthodes de bissection comportent deux inconvénients importants :

1. La fiabilité ainsi calculée nous fournit la précision des scores totaux obtenus pour la moitié du test, alors que c'est la précision des scores totaux pour l'ensemble du test qui nous intéresse. De plus, une telle estimation de la fiabilité risque de fournir des estimations bien au-dessous de la fiabilité du score total pour l'ensemble du test puisque plus un score total est calculé sur un grand nombre d'items, plus il a de chances d'être précis.
2. La fiabilité ainsi calculée dépend de la méthode de bissection choisie. En effet, les corrélations entre les scores obtenus aux deux moitiés d'un test risquent d'être fort différentes selon que les deux moitiés sont constituées des premiers vs derniers items, des items pairs vs items impairs ou encore des *j/2* items choisis au hasard vs *j/2* items restants.

Le tableau 5 présente deux exemples de calcul selon la méthode de bissection : la première utilise la corrélation entre les deux parties du test formées de la somme des items pairs et impairs ; la seconde utilise la corrélation entre la somme des cinq premiers et des cinq derniers items. L'estimation de la fiabilité varie beaucoup selon la méthode de bissection employée. Dans le premier cas, celle-ci est estimée à 0,973, alors que dans le second cas, elle est évaluée à 0,887. D'autres méthodes de bissection auraient fourni des résultats tout aussi différents.

La correction de Spearman-Brown permet d'apporter une solution pratique au problème de la sous-estimation de la fiabilité de l'ensemble du test par la méthode des deux moitiés. Cette formule de correction permet d'estimer quelle serait la fiabilité du test entier à partir de la fiabilité calculée entre deux moitiés. Cette correction prend la forme suivante :

$$\hat{r}_{XX'} = \frac{2r_{XX'}}{1 + r_{XX'}} \qquad (4.34)$$

où $r_{XX'}$ représente la corrélation entre les deux moitiés d'un test.

Dans le cas où la corrélation entre les deux moitiés d'un test serait de 0,81, la correction de Spearman-Brown estimerait la fiabilité du test entier à :

$$\hat{r}_{XX'} = \frac{2 \times 0,81}{1 + 0,81} = \frac{1,62}{1,81} = 0,90 \qquad (4.35)$$

Cette estimation n'est toutefois valable que si les deux moitiés du test correspondent à la définition de deux tests strictement parallèles. Lorsque les variances des deux moitiés sont fort différentes, l'estimation de la fiabilité du test entier risque d'être faussée.

Rulon (1939) a proposé une alternative à la méthode Spearman-Brown qui fournit une meilleure estimation de la fiabilité du test entier lorsqu'il y a de grandes différences dans les variances des scores calculés à partir des deux moitiés. La formule de Rulon suppose que l'on calcule d'abord un score de différence entre les résultats aux deux moitiés du test pour chaque sujet :

$$D = A - B \qquad (4.36)$$

La fiabilité du test entier est ensuite calculée à partir de la formule suivante :

$$r_{XX'} = 1 - \frac{s_D^2}{s_X^2} \qquad (4.37)$$

$r_{XX'}$ est la fiabilité du test entier, s_D^2 est la variance des scores de différence et s_X^2 est la variance des scores observés. Il existe également une autre formule de correction qui donne exactement les mêmes résultats que celle de Rulon, mais attribuable à Guttman (1945).

Le deuxième inconvénient des méthodes de bissection est plus sérieux. En effet, selon les deux moitiés obtenues par la méthode de bissection choisie, il y a autant d'estimations possibles de la fiabilité. La meilleure estimation de la cohérence interne serait obtenue en calculant la moyenne des estimations obtenues à partir de toutes les bissections possibles du test. Ceci représenterait, cependant, une quantité énorme de calculs même pour un test comportant relativement peu d'items.

3.3 MÉTHODE DES COVARIANCES

Les méthodes d'estimation de la cohérence interne fondée sur la covariance entre les items permettent d'apporter une solution au problème que nous venons de souligner. Ces méthodes reposent sur le postulat que chaque item peut être considéré comme une partie d'un test et qu'un test peut être considéré comme étant composé d'autant de parties que d'items. Plus les covariances entre tous les items pris deux à deux sont élevées, plus les items sont homogènes et mesurent la même chose. Ceci se traduit par un score total qui sera d'autant plus précis qu'il sera évalué par un échantillon d'items tirés de la même population. S'il n'y a que peu de covariances entre les items, alors de nouveaux échantillons tirés de la même population risquent de se traduire par des résultats au test fort différents. En fait, ce que font les méthodes de covariance, c'est d'estimer les chances d'obtenir le même résultat avec de nouveaux échantillons d'items, à partir des corrélations qui existent déjà entre items censés mesurer la même chose.

Le α de Cronbach est sans doute la méthode la plus connue d'estimation de la cohérence interne fondée sur les covariances entre items (Cronbach, 1951). C'est aussi l'une des plus utilisées. Plusieurs logiciels statistiques fournissent maintenant cette valeur de façon routinière. La valeur du α de Cronbach est estimée à partir de l'équation suivante, qui est le résultat du calcul de la fiabilité d'un score composite à partir de la fiabilité de ses parties (ou items) :

$$\alpha = \frac{j}{j-1}\left[1 - \frac{\sum s_j^2}{s_x^2}\right] \qquad (4.38)$$

Dans cette formule, j représente le nombre d'items, $\sum s_j^2$, la somme des variances des j items et, la variance des scores totaux au test. Le α de Cronbach repose sur le postulat fort que chaque item est parallèle aux autres (même degré de difficulté, même variance). Comme c'est rarement le cas dans la pratique, la valeur de fiabilité fournie

par α sous-estime la fiabilité du score total au test. On peut donc affirmer que α est une valeur conservatrice de la cohérence interne du score total puisque $\alpha \leq r_{xx}$.

Tableau 6 – Résultats et matrice des variances-covariances pour un test de 10 items

Sujets#	Items #									
	1	2	3	4	5	6	7	8	9	10
1	2	2	3	1	3	4	1	2	4	1
2	4	3	3	1	3	5	2	4	5	2
3	5	3	5	3	4	5	3	5	5	2
4	2	1	3	2	3	4	1	3	5	1
5	3	2	5	4	5	5	2	4	5	1
6	3	3	5	4	5	5	2	4	5	1
7	1	1	3	2	3	3	0	2	4	2
8	1	1	3	1	4	3	0	1	3	0
9	0	1	2	0	3	3	0	1	3	0
10	5	3	5	3	3	5	3	5	5	3
Moyenne	2,6	2,0	3,7	2,1	3,6	4,2	1,4	3,1	4,4	1,3
Variance	2,6	0,8	1,2	1,7	0,6	0,8	1,2	2,1	0,6	0,8

Matrice des variances-covariances

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Item 1	2,64									
Item 2	1,30	0,80								
Item 3	1,38	0,70	1,21							
Item 4	1,24	0,60	1,33	1,69						
Item 5	0,24	0,20	0,58	0,74	0,64					
Item 6	1,28	0,70	0,76	0,78	0,28	0,76				
Item 7	1,76	0,90	1,02	0,96	0,26	0,92	1,24			
Item 8	2,24	1,10	1,33	1,39	0,34	1,18	1,56	2,09		
Item 9	1,06	0,50	0,62	0,76	0,16	0,62	0,74	1,06	0,64	
Item 10	1,12	0,50	0,49	0,47	-0,18	0,44	0,68	0,97	0,48	0,81

Le tableau 6 présente la matrice des variances-covariances pour les 10 items d'un test. Les variances sont inscrites en diagonale, alors que les covariances entre items pris deux à deux figurent au-dessous de la diagonale principale.

La valeur du α de Cronbach de ce test est calculée de la manière suivante (S_x^2 = 87,6) :

$$\alpha = \frac{10}{9}\left[1 - \frac{12,4}{87,6}\right] = 0,95 \qquad (4.39)$$

Dans cet exemple, le nombre d'items j est égal à 10 et la somme des variances des items (diagonale de la matrice des variances-covariances) est égale à 12,4. La variance des scores totaux au test est de 87,6. Le α de Cronbach ainsi calculé est égal à 0,95, ce qui représente une excellente cohérence interne. Les dix items forment un ensemble suffisamment homogène pour qu'il soit justifié d'additionner ensemble leurs résultats pour former un score total.

L'inspection de la matrice des variances-covariances ne révèle qu'une covariance négative, entre l'item 5 et l'item 10. Il est à prévoir que ces items contribuent moins à la cohérence interne du test entier. Un item qui possède une covariance négative avec les autres items ne saurait être considéré comme faisant partie du même groupe d'items.

Il est difficile d'interpréter le degré d'association entre deux items à partir de la matrice des variances-covariances. Lorsque les items ont des étendues différentes, la valeur de la covariance s'en trouve affectée. Un item qui peut prendre les valeurs de 1 à 10 aura vraisemblablement une covariance plus élevée qu'un item qui ne peut prendre que des valeurs de 1 à 2 ou de 1 à 5. C'est pourquoi il est préférable d'avoir recours à la matrice des corrélations dans ce genre de situation. Le tableau 7 présente la matrice des corrélations entre items pour les mêmes données que celles figurant dans le tableau 6.

Lorsque l'étendue est la même pour tous les items, comme c'est le cas dans le tableau 6, la covariance s'interprète plus simplement. Plus celle-ci est élevée, plus elle indique une forte association entre les deux items. Il est également important de remarquer que la covariance entre deux items sera toujours inférieure à la plus petite des variances de chacun des deux items. En effet, comment la variance de l'item — la covariance de l'item avec lui-même — pourrait-elle être inférieure à la covariance de l'item avec un autre item.

Tableau 7 – Résultats et matrice des corrélations d'un test de 10 items

Sujets#	Items # 1	2	3	4	5	6	7	8	9	10
1	2	2	3	1	3	4	1	2	4	1
2	4	3	3	1	3	5	2	4	5	2
3	5	3	5	3	4	5	3	5	5	2
4	2	1	3	2	3	4	1	3	5	1
5	3	2	5	4	5	5	2	4	5	1
6	3	3	5	4	5	5	2	4	5	1
7	1	1	3	2	3	3	0	2	4	2
8	1	1	3	1	4	3	0	1	3	0
9	0	1	2	0	3	3	0	1	3	0
10	5	3	5	3	3	5	3	5	5	3
Moyenne	2,6	2	3,7	2,1	3,6	4,2	1,4	3,1	4,4	1,3
Variance	2,6	0,8	1,2	1,7	0,6	0,8	1,2	2,1	0,6	0,8

Matrice des variances-covariances

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item 6**	**Item 7**	**Item 8**	**Item 9**	**Item 10**
Item 1	1									
Item 2	0,895	1								
Item 3	0,772	0,711	1							
Item 4	0,587	0,516	0,930	1						
Item 5	0,185	0,280	0,659	0,712	1					
Item 6	0,904	0,898	0,793	0,680	0,401	1				
Item 7	0,973	0,904	0,833	0,663	0,292	0,948	1			
Item 8	0,954	0,851	0,836	0,740	0,294	0,936	0,969	1		
Item 9	0,815	0,699	0,705	0,731	0,250	0,889	0,831	0,917	1	
Item 10	0,766	0,621	0,495	0,402	-0,250	0,561	0,679	0,746	0,667	1

Parmi les autres méthodes d'estimation de la cohérence interne du score total d'un test, il y a les formules 20 et 21 développées par Kuder et Richardson (1937). La formule 20 permet de calculer la cohérence interne pour des items dichotomiques, alors que la formule 21 permet d'effectuer les mêmes calculs à partir de la moyenne et de la variance des scores individuels. Lorsque tous les items ont sensiblement la même difficulté et la même variance, les deux formules fournissent des estimations équivalentes. Cependant, lorsque les items varient beaucoup en difficulté et en variance, la formule 21 fournit une estimation de la cohérence interne systématiquement inférieure.

La formule 20 de Kuder-Richardson est la suivante :

$$KR_{20} = \frac{j}{j-1}\left[1 - \frac{\sum pq}{s_x^2}\right] \tag{4.40}$$

Dans l'équation (4.39), j et ont la même signification que dans l'équation 4.37. La seule différence importante provient de l'expression (4.40) :

$$\sum pq \tag{4.41}$$

qui sert au calcul de la somme des variances des items lorsque ceux-ci sont corrigés de façon dichotomique. Dans ce cas, p est le coefficient de difficulté de l'item et $q=1-p$.

Lorsque les items sont tous sensiblement de même difficulté et de même variance, la formule 20 peut être remplacée par l'approximation suivante fournie par la formule 21 :

$$KR_{21} = \frac{j}{j-1}\left[1 - \frac{\bar{X} - (j - \bar{X})}{js_x^2}\right] \tag{4.42}$$

La formule 21 permet d'estimer la cohérence interne d'un test à partir de la moyenne $\bar{X}$ et de la variance des scores totaux s_x^2. Cependant, s'il y a d'importantes différences parmi les indices de difficulté des items, KR21 sera systématiquement inférieure à KR20.

Enfin, Hoyt (1941) a mis au point une méthode de calcul de la cohérence interne qui fournit des résultats similaires à la valeur α de Cronbach en utilisant cette fois-ci le modèle de l'analyse de variance en blocs aléatoires. Hoyt définit la fiabilité de cohérence interne de la manière suivante :

$$r_{XX'} = \frac{MC_{personnes} - MC_{erreur}}{MC_{personnes}} \tag{4.43}$$

Dans cette expression, $MC_{personnes}$ représente la moyenne des carrés des personnes (ou si l'on préfère la variance des scores observés) et MC_{erreur} représente la moyenne des carrés d'erreur (ou, si l'on préfère, la variance d'erreur aléatoire). La différence entre les deux termes du numérateur permet d'estimer la variance des scores vrais, soit la variance des scores observés qui n'est pas de l'erreur. Le rapport entre la variance des scores vrais et la variance des scores observés nous fournit l'expression habituelle de la fiabilité.

Afin de bien comprendre la formule de Hoyt, il faut tenir compte que ces moyennes de carré sont calculées à partir d'un modèle d'analyse de variance en blocs aléatoires. Dans ce modèle, il n'y a pas de terme d'interaction : la variance d'interaction est en effet confondue avec la variance d'erreur. L'absence d'interaction signifie que, selon ce modèle, la difficulté des items est la même pour chaque personne ayant répondu au test.

Par exemple, l'item le plus difficile a été le plus difficile pour tous les sujets et non pour certains d'entre eux. Si cela devait être le cas et qu'il devait y avoir une interaction significative entre la difficulté des items et la personne qui y répond, alors cette variance s'ajouterait à la variance d'erreur et contribuerait à réduire la fiabilité des résultats au test : un item deviendrait facile ou difficile selon la personne qui y répond. Dans un telle situation, il est difficile d'espérer une quelconque fiabilité des résultats dans l'appréciation des différences individuelles.

Le tableau 8 présente les résultats de l'analyse de variance effectuée sur les données du tableau 6. On y retrouve les sources de variance, la somme des carrés et la moyenne des carrés pour un plan en blocs aléatoires. En appliquant la formule de Hoyt (équation 4.43) aux résultats de ce tableau, on retrouve la même valeur de α que celle calculée par l'équation (4.38).

$$\alpha = \frac{9{,}74 - 0{,}46}{9{,}74} = 0{,}95 \qquad (4.44)$$

Tableau 8 – Analyse de variance des résultats à un test de 10 items

Source de variance	Somme des carrés	Degrés de liberté	Moyenne des carrés
Personnes	87,64	9	9,74
Items	151,80	90	1,69
Erreur	37,56	81	0,46
Total	239,44	99	2,42

La formule de Hoyt anticipe sur les développements futurs apportés par la théorie de la généralisabilité telle que formulée par Cronbach, Glaser, Nanda et Rajaratnam (1972). Grâce à l'étude des composantes de variance, la théorie de la généralisabilité, comme nous le verrons plus loin, permet l'étude de la fiabilité des scores dans des conditions d'observation beaucoup plus complexes que celles que nous avons considérées dans ce chapitre. Enfin, grâce aux modifications apportées par Cardinet et Tourneur (1985), la théorie de la généralisabilité a permis d'étendre l'étude de la fiabilité de manière à y inclure tout objet de mesure dont les niveaux sont échantillonnés aléatoirement, que ce soit la difficulté des items eux-mêmes ou des items à l'intérieur d'objectifs d'apprentissage, par exemple.

4. Facteurs affectant l'estimation de la fiabilité des résultats

Les facteurs affectant l'estimation de la fiabilité des résultats à un test proviennent de deux sources principales :

- les limites inhérentes au calcul de la corrélation linéaire au moyen du *r* de Pearson ;
- les conditions empiriques de l'administration du test, tels que la longueur du test et la limite de temps imposée.

Parce que, dans la pratique, l'estimation de la fiabilité procède par un calcul de corrélation, les valeurs de fiabilité dépendent du modèle de la corrélation linéaire de Pearson et des postulats de ce type de calcul statistique (voir chapitre 2). Les limites statistiques du *r* de Pearson s'étendent donc au coefficient de fiabilité. Voici un bref rappel de ces limites dont il faut tenir compte dans toute interprétation d'un coefficient de fiabilité.

4.1 La difficulté d'un test

Celle-ci affectera le calcul de la fiabilité parce qu'un test trop facile ou trop difficile entraînera une certaine asymétrie des résultats : asymétrie positive dans le cas d'un test trop difficile, asymétrie négative dans le cas d'un test trop facile. Or, la corrélation r de Pearson ne peut atteindre sa valeur maximum de 1 que lorsque les distributions des deux variables en corrélation sont symétriques ou possèdent le même type d'asymétrie.

Prenons le cas du calcul d'un coefficient de stabilité au moyen de la corrélation test-retest. Dans la situation où les scores se distribuent de manière symétrique lors d'une première administration, puis de manière asymétrique lors d'une seconde administration, la valeur maximale du coefficient de corrélation entre les scores au test et au retest ne pourra atteindre la valeur maximum de +1.

Il est donc important de prendre en considération les facteurs affectant la fiabilité. Dans ce dernier cas, il est tout aussi important — sinon plus — de savoir que la distribution des scores a changé que de savoir que la valeur de stabilité est faible. En effet, le changement de distribution peut expliquer pourquoi la fiabilité est faible. Un test devenu trop facile au moment du retest peut expliquer que la distribution des résultats, symétrique au moment du test, soit devenue asymétrique négative au moment du retest. La contamination des résultats ou l'apprentissage peuvent expliquer ce genre de phénomène.

4.2 L'étendue des différences individuelles

La variance totale d'un test est une condition nécessaire mais non suffisante à la fiabilité des résultats. C'est ce que nous avons vu en traitant de la variance du score total à un test. Toute réduction de l'étendue des scores individuels entraîne une sous-estimation de la corrélation entre deux variables (voir chapitre 2).

Lors de l'étude de la fiabilité d'un instrument de mesure, plusieurs situations peuvent se produire contribuant à réduire les différences individuelles et, par consé-

quent, nos chances d'obtenir une estimation correcte de la fiabilité. C'est le cas, notamment, des situations suivantes :

1. L'étude-pilote porte sur un échantillon qui possède une variance moindre que la population générale. C'est le cas d'un test dont les résultats ne sont recueillis que dans des écoles provenant de milieux favorisés. On peut suspecter que la variance des résultats ainsi recueillis est moindre que celle qui aurait été obtenue au moyen d'un échantillon représentatif.
2. Un test a été mis à l'essai sur une population scolaire à plusieurs niveaux, plus étendue que le seul niveau dans lequel le test doit être employé. Il faut être prudent dans l'appréciation de la fiabilité rapportée par les auteurs de ce test. Les résultats peuvent donner lieu à une variance des scores qui soit artificiellement grande lorsque les répondants sont de plusieurs niveaux scolaires. Par contre, cette variance risque d'être réduite, et la fiabilité de même, si l'on emploie le test à un seul niveau scolaire.

Magnusson (1967) a mis au point une formule permettant de corriger l'estimation de la fiabilité lorsque nous avons de bonnes raisons de croire que notre échantillon de sujets est homogène et contribue ainsi à sous-estimer la variance totale des scores observés au test. Cette formule de correction est donnée par l'équation suivante :

$$r_{UU'} = 1 - \frac{S_x^2(1 - r_{XX'})}{S_U^{\ 2}} \qquad (4.45)$$

Dans cette équation, $r_{XX'}$ est la fiabilité estimée pour le nouvel échantillon, S_x^2 est la variance de l'échantillon original, S_u^2 est la variance du nouvel échantillon et $r_{XX'}$ est la fiabilité estimée à partir de l'échantillon de départ.

Cette correction de Magnusson postule que l'erreur aléatoire est la même dans les deux groupes et que la différence dans les variances des scores observés est imputable à des différences dans les variances des scores vrais dans les deux groupes. C'est pourquoi, lors de l'utilisation de normes, il est important de s'assurer que notre échantillon provient de la même population qui a servi au calcul des valeurs de la fiabilité du test, sinon il sera plus prudent de réaliser une étude-pilote sur la fiabilité des résultats obtenus avec l'échantillon concerné.

4.3 Limite de temps

Lorsqu'un test est chronométré, plusieurs élèves n'arrivent pas à répondre à toutes les questions. Les questions omises se trouvent généralement à la fin du test et celles-ci sont généralement cotées 0. Cette procédure a pour effet de créer une inflation artificielle de la corrélation entre les derniers items, ce qui aura pour effet de faire paraître ces items plus homogènes qu'ils ne le sont en réalité. Cette homogénéité ne sera pas due au fait que les items mesurent la même chose, mais plutôt au fait qu'ils ont été omis par les sujets parce qu'ils se trouvaient en fin de test.

Il faut donc être très prudent lorsque l'on administre un test chronométré et que l'on souhaite déterminer la fiabilité des résultats. L'estimation de la fiabilité risque d'être faussée par la corrélation artificielle entre les items dans le cas des méthodes de bissection ou encore de cohérence interne (α de Cronbach). L'usage de la méthode test-retest sera alors préférable.

4.4 La longueur du test

Plus un test comprend un grand nombre d'items correspondant à ce que nous souhaitons mesurer, plus cette mesure devrait être précise. En effet, la somme des erreurs aléatoires de mesure devrait tendre vers zéro lorsqu'un grand nombre d'items est utilisé. C'est le principe de la théorie de l'échantillonnage : plus un échantillon est grand, plus l'estimation des caractéristiques de la population dont il est tiré tend à être précise.

Le rapport entre la longueur d'un test et sa fiabilité est exprimée par la formule de Spearman Brown (*Spearman Brown prophecy formula*). Elle nous indique à quel degré de précision l'on peut s'attendre d'un test dont l'on modifierait le nombre d'items dans une proportion k (k pouvant être une fraction ou un entier). Voici un rappel de cette formule que nous avons déjà vue dans le cas de la méthode de bissection où $k = 2$ (formule 4.34) :

$$\hat{r}_{XX'} = \frac{k r_{YY'}}{1 + (k-1) r_{YY'}} \tag{4.46}$$

Dans l'équation précédente, $\hat{r}_{XX'}$ représente la fiabilité attendue du test modifié, $_{YY'}$ représente la fiabilité du test initial. Lorsque $k > 1$, nous calculons la fidélité pour un test allongé. Par exemple, si un test comporte 12 items et que l'on souhaite connaître la fidélité de ce test auquel nous avons ajouté 18 items parallèles, soit 30 items en tout, alors nous utilisons la formule (4.46) avec $k = 2{,}5$ (2,5 x 12 = 30). Le même principe s'applique pour $k < 1$. Les valeurs de fiabilité calculées le sont alors pour des tests plus courts.

La formule de Spearman-Brown nous permet de déterminer dans quelle proportion la longueur d'un test doit être augmentée pour atteindre un degré visé de fiabilité. En modifiant l'équation précédente, l'on peut isoler k de la façon suivante :

$$k = \frac{r_{XX'}(1 - r_{YY'})}{r_{YY'}(1 - r_{XX'})} \tag{4.47}$$

Supposons que l'on veuille estimer dans quelle proportion un test de 30 items doit être prolongé pour que sa fiabilité, actuellement de 0,75, soit portée à 0,85. En solutionnant l'équation (4.47) pour trouver k, on obtient :

$$k = \frac{0{,}85\ (1 - 0{,}75\ )}{0{,}75\ (1 - 0{,}85\ )} = 1{,}89$$

Une valeur k = 1,89 signifie que le nouveau test devra être 1,89 fois plus long que le test original. Il devra donc compter approximativement 1,898 x 30 items, soit 57 items. Il faudrait donc ajouter 27 items aux 30 items faisant déjà partie du test pour faire passer la fiabilité du test de 0,75 à 0,85.

Il est important de se rappeler que la formule de Spearman Brown prend pour acquis que les items qui seront ajoutés (ou retranchés) sont parallèles aux items du test de départ, c'est-à-dire qu'ils sont de même contenu et de même degré de difficulté. En effet, la précision d'un test n'augmentera pas si l'on y ajoute des items de difficulté ou

de contenu différent, susceptibles de ne pas avoir une bonne corrélation avec les items faisant déjà partie du test.

La formule de Spearman Brown peut être très utile pour nous permettre de décider de la longueur qu'un test doit avoir pour posséder une précision acceptable. Cependant, cette méthode ne nous indique pas quelles sont les caractéristiques des items parallèles à ajouter, en termes de contenu et de format, afin d'accroître la fiabilité des tests. Lorsque le contenu d'un test est défini de façon générale, comme c'est le cas de plusieurs épreuves sommatives, le constructeur peut avoir de la difficulté à définir les caractéristiques des items à ajouter pour qu'ils soient des items parallèles à ceux déjà construits. En éducation, par exemple, le concepteur pourra s'inspirer des objectifs pédagogiques pour ajouter des items provenant des mêmes objectifs que le test initial. Plus les conditions ayant présidé à l'élaboration initiale du test sont claires — comme c'est le cas avec les techniques de spécification de domaine, plus il sera facile au concepteur de rédiger des items parallèles.

Le principal inconvénient de cette manière de procéder est de postuler que tous les items possèdent sensiblement la même homogénéité. Il est possible que certains items possèdent des caractéristiques qui leur permettent de mesurer de façon plus précise les sujets d'un échantillon particulier. Il est plus facile d'améliorer la fiabilité d'un test lorsque celui-ci a été construit selon des facettes ou une approche critériée (voir chapitre 3).

5. Fiabilité et erreur de mesure

La fiabilité n'exprime pas la précision d'une mesure dans le même système d'unités que le score total ce qui en rend l'interprétation difficile. C'est pourquoi, plutôt que de rapporter la précision d'un test sous forme de fiabilité, on préfère parfois indiquer l'erreur qui entoure l'interprétation d'un score. Plus les résultats à un test sont fiables, plus l'erreur entourant un score sera faible.

Dans la pratique, il existe deux façons de calculer l'intervalle de confiance entourant le score observé de l'individu. On peut postuler que plus le test est fiable, plus cet intervalle sera petit, car meilleure sera l'estimation du score vrai par le score observé de l'individu. Voici deux occasions où cette situation se présente :

1. On est intéressé à déterminer l'intervalle de confiance autour du score observé à l'intérieur duquel se situe le score vrai de l'individu : *l'erreur de mesure.*
2. On est intéressé à déterminer l'intervalle de confiance du score observé d'un élève s'il devait être soumis à un test parallèle au premier : *l'erreur d'estimation.*

5.1 L'ERREUR TYPE DE MESURE

Pour comprendre cette notion, nous devons nous rappeler que, dans la théorie de la note vraie, les scores d'un individu se distribuent normalement autour d'une valeur moyenne qui correspond à sa note vraie. Nous pouvons calculer l'écart type de cette distribution. Si nous faisons de même pour tous les sujets d'un groupe donné, nous pourrons calculer la moyenne des écarts types des différentes distributions. Cet

écart type moyen est appelé l'erreur type de mesure (S_E). Elle peut être estimée grâce à la formule suivante :

$$S_E = S_X\sqrt{1 - r_{XX'}} \quad (4.48)$$

$r_{XX'}$ = le coefficient de fiabilité du test

S_X = l'écart type de la distribution des résultats sur lesquels a été calculé

Par exemple, si S_X $r_{XX'}$ est égal à 0,90 et S_X est égal à 15 alors $S_E = 15\sqrt{1 - 0,90}$ = 4,75 ≃ 5. Mais que signifie pratiquement une erreur type de 5 ? Partant du postulat que l'erreur de mesure se distribue normalement, nous pouvons nous attendre à ce que, pour un sujet donné, 68% de ses scores observés tombent dans un intervalle de ± 1 écart type autour de son score vrai (figure 6). Par conséquent, dans notre exemple, en supposant que le score vrai est égal à 110, nous avons 2 chances sur 3 (rapport de 68 et 32) d'observer, lors d'une passation quelconque, une note comprise entre 105 et 115 (= 110 ± 5). Si nous voulons une probabilité plus grande, il nous faudra élargir notre intervalle. Ainsi, si nous voulons avoir 99% de chance (2,58 écarts types) que la note obtenue tombe dans un intervalle déterminé, nous devrons définir un intervalle de 13 points (13 = 2,58 x 5) de part et d'autre de la moyenne.

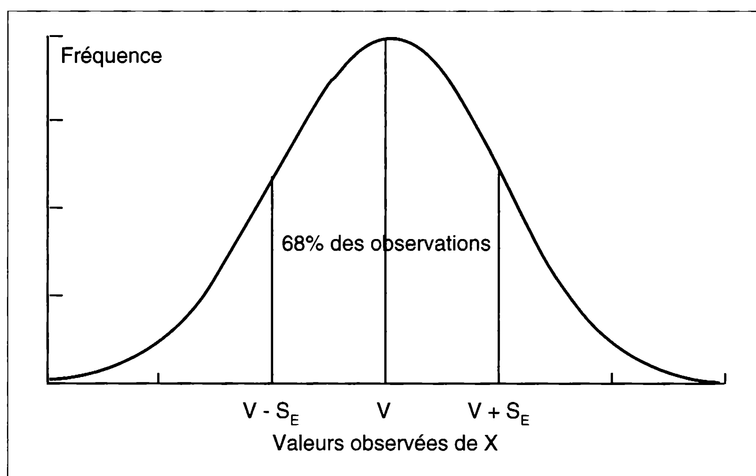


Figure 6 – Distribution attendue des scores d'un sujet pour un testing répété

Mais, dans la pratique, nous ne connaissons évidemment pas le score vrai. Nous ne savons donc pas où se situe la note obtenue au sein de la distribution attendue des scores. Il se peut, par hasard, qu'elle soit égale au score vrai. Il se peut aussi qu'elle tombe à l'extrémité de la distribution. Mais cela, nous n'en savons rien. Par contre, nous connaissons l'erreur type de mesure et nous acceptons un risque d'erreur déterminé. À l'aide de ces informations, nous pouvons construire un intervalle de confiance autour de la note observée dans lequel le score vrai du sujet a un certain pourcentage de chance de se trouver. Si, dans notre exemple, le score obtenu par le sujet est de 104 points et que nous souhaitons déterminer un intervalle où le score vrai de ce sujet a 99% de chance de se trouver, nous allons construire un intervalle de ± 13 points autour

de 104. Cet intervalle sera donc égal à [91 ; 117]. Pour élargir ou pour rétrécir cet intervalle, il nous suffit de multiplier l'erreur type de mesure par la valeur critique de z correspondant au niveau de probabilité souhaité. Nous pouvons dès lors exprimer l'intervalle de confiance sous la forme générale suivante :

$$X - z_c s_E \leqslant V \leqslant X + z_c s_E \qquad (4.49)$$

X = le score observé,

z_c = la valeur critique de z,

s_E = l'erreur type de mesure,

V = le score vrai.

L'avantage majeur à déterminer un intervalle de confiance autour de la note obtenue est de relativiser cette dernière note. Le praticien prend ainsi mieux conscience de la marge d'erreur que comporte la mesure recueillie. Un large intervalle de confiance montre clairement que les scores observés ne sont pas de très bons indicateurs du score vrai. Inversement, un intervalle de confiance étroit est l'indice que les scores observés sont assez proches du score vrai.

D'un autre côté, l'intervalle de confiance a comme désavantage d'être parfois mal interprété par les praticiens. En fait, nous n'avons jamais de certitude que le score vrai soit inclus dans l'intervalle que nous avons établi autour de la note observée. Nous n'avons qu'une probabilité, plus ou moins importante selon le risque d'erreur choisi. Une autre limite de l'usage de l'intervalle de confiance est de postuler que l'erreur type de mesure est la même à tous les niveaux de performance (postulat d'homoscédasticité). Par exemple, dans le cas d'un test d'acquis scolaires, on suppose que l'erreur de mesure est la même pour les élèves brillants que pour les élèves faibles. La pertinence de ce postulat est discutable. Il se peut en effet que l'importance de l'erreur type de mesure diffère selon le niveau d'aptitude des sujets. Nous verrons plus loin que le modèle binomial des scores présente une méthode d'estimation de l'erreur type de mesure qui ne s'appuie pas sur le postulat d'homoscédasticité.

5.2 L'ERREUR TYPE D'ESTIMATION

Le deuxième type d'erreur est l'*erreur d'estimation*. C'est le même type d'erreur que l'on retrouve chaque fois que l'on souhaite calculer l'intervalle de confiance d'une valeur prédite à partir d'une équation de régression linéaire. Dans ce cas-ci, on cherche à prédire le résultat à un test parallèle à partir du résultat à un autre test parallèle. L'erreur type d'estimation est donnée par la racine carrée de la variance résiduelle. La variance résiduelle est la variance des scores qui subsiste dans le second test une fois que l'on tient compte des résultats au premier test. En effet, si l'on devait calculer la variance des scores observés au deuxième test à partir de tous ceux qui ont obtenu le même résultat (disons 15/20) au premier test, celle-ci devrait être d'autant plus petite que la corrélation (ou si l'on veut l'équivalence) entre les deux tests est grande. Plus les deux tests sont parallèles, plus grande est la probabilité que le score à l'autre test soit aussi de 15/20 ou une valeur très approchée. Un exposé détaillé du calcul et l'interprétation de l'erreur d'estimation a déjà été présenté au chapitre 2.

Rappelons que l'homoscédascité est souvent postulée lorsque l'on calcule des corrélations linéaires. Généralement, le chercheur intéressé à différencier les sujets entre eux suppose que l'erreur type d'estimation est la même pour tous. L'homoscédasticité rend plus simple l'interprétation des résultats. Néanmoins, le postulat d'homoscédasticité peut ne pas être réaliste dans d'autres situations. Par exemple, lorsque le chercheur veut estimer l'erreur de mesure qui entoure la proportion des items d'un domaine qu'un sujet est capable de réussir correctement. Il est naturel, dans une telle situation, que l'erreur de mesure soit moindre chez ceux qui réussissent ou échouent presque tous les items que chez ceux qui n'en réussissent que la moitié. Keats et Lord (1962) et Lord (1965) ont proposé un modèle de test fondé sur la distribution binomiale qui permet d'estimer ce type d'erreur de mesure particulièrement utile en mesure critériée. Nous aborderons ce modèle dans la section 6 de ce chapitre.

5.3 L'ERREUR TYPE DE LA DIFFÉRENCE

En dehors de la détermination d'un intervalle de confiance, la connaissance de l'erreur type de mesure est également utile si nous désirons comparer les scores obtenus par un même sujet à deux tests différents. Il est, par exemple, assez fréquent qu'un psychologue scolaire, au vu des résultats de tests, se demande si un sujet possède de meilleures aptitudes dans un domaine que dans un autre. Les écarts observés entre les scores à différents tests peuvent toutefois résulter de fluctuations aléatoires dues aux erreurs cumulées de chacune des mesures. Il est donc important de pouvoir estimer quelle est la probabilité qu'une différence observée soit le reflet d'une différence réelle entre les aptitudes d'un sujet. Dans ce but, nous pouvons calculer l'erreur type de la différence (s_{E_D}). Celle-ci est logiquement plus importante que l'erreur type de mesure de chacun des deux scores entre lesquels la différence est calculée.

Puisque, dans la théorie classique des scores, les erreurs de mesures sont non corrélées, nous pouvons écrire :

$$s^2_{E_D} = s^2_{E_X} + s^2_{E_Y} \tag{4.50}$$

$s^2_{E_D}$ = la variance de l'erreur de la différence,

$s^2_{E_X}$ et $s^2_{E_Y}$ = la variance de l'erreur des notes X et Y.

Pour être comparées, deux notes doivent être exprimées sur des échelles semblables. Autrement dit, les deux tests dont elles sont issues doivent avoir une même moyenne et un même écart type. Si c'est le cas, de l'équation ci-dessus, nous pouvons dériver l'équation suivante :

$$s^2_{E_D} = s_X\sqrt{2 - r_{XX'} - r_{YY'}} \tag{4.51}$$

s_{E_D} = l'erreur type de la différence,

s_X = l'écart type de la distribution des X, égal à celui de la distribution des Y,

r_{XX}' et r_{YY}' = le coefficient de fiabilité de chacun des deux tests.

Par exemple, si $r_{XX'} = 0{,}85$ et $r_{YY'} = 0{,}85$ et si $S_X = 10$ pour les deux tests, alors $S_{E_D} = 10\sqrt{2-0{,}88-0{,}85} = 5{,}20$. Cela signifie que la différence entre les deux tests doit être d'au moins 5,2 points pour être considérée comme statistiquement significative, avec un risque de se tromper dans 32% des cas. Si nous souhaitons un risque d'erreur moindre, nous devons multiplier l'erreur type de la différence par la valeur critique de z correspondant au seuil choisi. Dans notre exemple, si nous désirons réduire ce risque à 5% des cas, nous devrons multiplier 5,2 par 1,96, ce qui est égal à une différence 10,19 points. Cela signifie qu'une différence de 10 points ou plus n'a que 5% de chance de se produire du seul fait des fluctuations aléatoires dues aux erreurs de mesure. Il est raisonnable de considérer une telle différence comme statistiquement significative et de rejeter l'hypothèse d'une égalité des scores aux deux tests considérés.

6. Le modèle binomial de l'erreur

Les *Standards for Educational and Psychological Testing* (American Psychological Association, 1985, p. 22) recommandent aux constructeurs de test de communiquer l'erreur type de mesure pour différents niveaux de scores. Comme nous l'avons vu dans la section précédente, le théorie classique ne nous permet malheureusement pas de produire une telle information. La théorie classique s'appuie en effet sur le postulat d'une indépendance entre le score vrai et l'erreur de mesure. L'erreur type de mesure est dès lors estimée pour le test dans son ensemble, quel que soit le niveau de score vrai des sujets. Pour satisfaire à la recommandation des *Standards for Educational and Psychological Testing,* il est donc nécessaire de modifier certains postulats de la théorie classique afin de permettre une estimation de l'erreur type de mesure conditionnelle.

Il ne s'agit pas de se plier à une exigence purement formelle. La nécessité d'estimer l'erreur type de mesure à divers niveaux de scores est dictée par un certain nombre de constats empiriques. Ainsi, comparant plusieurs méthodes d'estimation de l'erreur type de mesure conditionnelle, Felt, Steffen & Gupta (1985, p. 358) observent que « *quelle que soit la méthode utilisée, l'erreur maximale est souvent deux fois plus importante que l'erreur minimale. Par conséquent, l'erreur type de mesure calculée selon la formule traditionnelle pour l'ensemble du test ne rend pas correctement compte de l'importance de l'erreur de mesure de beaucoup - et peut-être de la plupart - des sujets* ». Ce problème est particulièrement crucial dans le cas de tests critériés. Dans de tels tests, des valeurs seuils sont définies pour permettre de ranger les sujets dans différentes catégories comme, par exemple, la maîtrise ou la non maîtrise d'un apprentissage. La connaissance précise de l'erreur type de mesure pour chacun des score seuils est essentielle vu l'importance des décisions prises sur cette base. L'usage d'une unique erreur type de mesure pour l'ensemble des scores possibles au test risque en effet de conduire à des décisions inadéquates.

Le modèle binomial de l'erreur, développé par Lord (1955), permet de surmonter les limites de la théorie classique et de calculer des erreurs types de mesure conditionnelles, c'est-à-dire en fonction du niveau de score des sujets. Ce modèle n'est toutefois applicable que pour des items dichotomiques, c'est-à-dire cotés 1 ou 0. Dans

le cadre du modèle binomial de l'erreur, un test composé de n items dichotomiques est conçu comme un échantillon d'items tirés au hasard hors d'un univers d'items. Tous les items de cet ensemble sont réputés posséder les mêmes propriétés du point de vue du contenu, de la difficulté et de la discrimination. Cette situation est analogue à celle, classique en statistique, du tirage de boules dans une urne. Chaque sujet de la population est considéré comme capable de répondre correctement à une certaine proportion de l'ensemble des items. Cette proportion peut être conçue comme le nombre de boules blanches dans l'urne. Inversement, la proportion d'items auxquels le sujet est incapable de répondre correctement correspondrait au nombre de boules noires dans cette même urne.

Le score vrai d'un sujet est égal à la proportion de l'ensemble des items auxquels il peut répondre correctement. Dans les faits, le sujet ne répond qu'à un test particulier, c'est-à-dire à un échantillon d'items tirés aléatoirement de l'ensemble des items. Si nous constituons aléatoirement un très grand nombre de tests à partir de cet ensemble d'items, la distribution des scores d'un sujet à tous ces tests se distribuera autour du score vrai de ce sujet. L'erreur type de mesure sera alors égale à l'écart type de cette distribution. Mais comment estimer cette erreur type lorsque nous disposons seulement du score du sujet à un seul test ? Pour répondre à cette question, nous devons nous souvenir que les items sont tous dichotomiques. Par conséquent, la distribution de fréquence des scores aux différents tests constitués aléatoirement à partir d'un vaste ensemble d'items correspondra approximativement à la distribution binomiale. Rappelons que, mathématiquement, la distribution binomiale est définie par la formule suivante :

$$P(X) = \frac{N!}{X!\,(N-X)!}p^{X}q^{(N-X)} \qquad (4.52)$$

$P(X)$ = la probabilité de X succès,

N = le nombre de tirages,

$N!$ = factoriel N = le produit de tous les entiers de N jusque 1
= $N(N-1)(N-2)(N-3)\dots 1$,

p = la probabilité de succès lors d'un tirage quelconque,

q = $(1-p)$ = la probabilité d'échec lors d'un tirage quelconque.

Supposons qu'un sujet soit capable de répondre correctement à 75% des items et qu'il ait à passer un test de 12 items. Son score vrai est donc égal à 9 (= 75% de 12). Toutefois, son score observé peut fluctuer aléatoirement autour de cette valeur du fait de l'erreur de mesure. Grâce à la formule (4.52), nous pouvons estimer la probabilité que ce sujet obtienne un score donné, différent de 9. Calculons, par exemple, la probabilité que ce sujet réponde correctement à 11 items, c'est-à-dire que son score total soit égal à 11 puisque les items sont cotés 1 ou 0 ? Appliquons la formule (4.52) :

$$P(11) = \frac{12!}{11!\,(12-11)!} \times 0{,}75^{11} \times 0{,}25^{(12-11)} = 0{,}1267 \qquad (4.53)$$

Ce résultat signifie que, si un sujet possède la capacité de réussir 75% des items et qu'il doit passer un test de 12 items constitué de manière aléatoire, il obtiendra un score de 11 points lors d'un peu plus de 12 passations sur 100. Nous pouvons, de la

même manière, calculer la probabilité que ce sujet obtienne chacun des 13 scores possibles à un test de 12 items (0 est un des scores possibles). Les probabilités que nous obtiendrons nous permettrons de déterminer la distribution de fréquence des scores attendus à un test de 12 items pour un sujet dont le score vrai est égal à 9.

Lorsque le nombre d'items est supérieur à 30, la loi normale constitue une bonne approximation de la distribution binomiale. Pour le calcul des probabilités associées à un score particulier, on peut transformer le nombre d'items réussis en score z et trouver sa probabilité dans la table de probabilités de la loi normale. Les caractéristiques de cette distribution peuvent être calculées à l'aide des formules suivantes :

$$\mu = Np \tag{4.54}$$

$$\sigma^2 = Npq \tag{4.55}$$

$$\sigma = \sqrt{Npq} \tag{4.56}$$

Supposons que notre test soit composé de 30 items. Un élève en a réussi 80%, soit 24. Le score moyen obtenu par le sujet sera donc égal à 24, c'est-à-dire à son score vrai. La variance des scores sera, elle, égale à 4,8 et l'écart type égal à 2,19. Nous avons vu plus haut que cet écart type correspond en fait à l'erreur type de mesure. Cela signifie que, pour une score vrai de 24 points, nous avons un peu plus de 68 chances sur 100 d'observer, lors d'une passation de test quelconque, un score inclus dans l'intervalle de ± 2,19 points autour de 24.

Dans la pratique, nous ne connaissons évidemment pas le score vrai du sujet que nous évaluons. Pour calculer l'erreur type de mesure, nous devons alors prendre la proportion d'items réussis par ce sujets comme estimation de son score vrai. Par ailleurs, il est également nécessaire d'introduire dans la formule (4.56) une correction pour obtenir une estimation non biaisée de la variance de la population. Nous obtenons alors la formule nous permettant d'estimer l'erreur type de mesure pour un score observé donné :

$$S_E = \sqrt{Npq\left(\frac{N}{N-1}\right)} \tag{4.57}$$

N = nombre d'items du test,

p = proportion d'items réussis = score total au test divisé par N,

q = (1 - p).

Par exemple, nous pouvons calculer l'erreur type de mesure d'un score de 6 points à un test homogène de 12 items :

$$S_E = \sqrt{12 \times 0{,}5 \times (1 - 0{,}5) \times \left(\frac{12}{12-1}\right)} = (2{,}75)$$

Si nous réalisons le même calcul pour chacun des scores possibles à ce test de 12 items, nous pouvons constater que l'erreur type de mesure est maximale au centre de la distribution des scores. Elle est par contre minimale à chacune des extrémités de cette même distribution. Nous pouvons ainsi constater que, contrairement au troisième postulat de la théorie classique, l'erreur de mesure peut être différente selon le score

vrai. Le modèle binomial de l'erreur nous permet de tenir compte de ces changements. Toutefois, cette amélioration par rapport à la théorie classique se fait au prix de postulats plus exigeants, ce qui conduit Lord (1965) à qualifier le modèle binomial de théorie forte du score vrai *(strong true-score theory)*. Deux postulats doivent, en particulier, retenir notre attention. Le premier concerne l'indépendance locale des items. Cela signifie qu'à un niveau de score vrai donné, les résultats à chaque item doivent être indépendants les uns des autres. Un second postulat est qu'à un niveau de score vrai donné, la probabilité de réussite est identique pour tous les items de l'ensemble considéré. Ce dernier postulat est, en pratique, quasi impossible à satisfaire. Pour prendre en compte les inévitables variations de difficulté, Keats (1957) a proposé une version sensiblement modifiée de la formule (4.57) :

$$S_E = \sqrt{Npq\left(\frac{N}{N-1}\right)\left(\frac{1-r_{XX'}}{1-r_{21}}\right)} \tag{4.58}$$

$r_{XX'}$ = le coefficient de fiabilité (formes parallèles, bissection ou alpha),

r_{21} = la formule 21 de Kuder-Richardson (4.42)

En fait, la formule (4.58) est identique à la formule (4.59) hormis l'introduction d'un facteur de correction qui a pour effet de réduire, en moyenne, les estimations des erreurs types de mesure et de les ramener à un niveau plus adéquat. Felt et al. (1985) recommandent l'utilisation de cette formule de préférence à la formule (4.57). Lord (1965) a proposé une modification du modèle binomial de l'erreur pour tenir compte du fait que de nombreux tests incluent des items de différents niveaux de difficulté. Dans le *modèle binomial composite de l'erreur*, on conçoit les formes parallèles d'un test comme des échantillons stratifiés d'items plutôt que comme des échantillons simplement aléatoires. En d'autres termes, au lieu de tirer les boules d'une même urne, nous les tirons de plusieurs urnes qui, chacune, contiennent une proportion différente de boules blanches. Aux urnes correspondent des ensembles d'items dont le niveau de difficulté diffère. Chacun des ces ensemble constitue une strate. Nous devons prendre en compte autant de strates qu'il y a de niveaux de difficulté au sein du test considéré. Ce genre de situation se rencontre en éducation dans les tests de maîtrise centrés sur plusieurs objectifs. L'erreur type de mesure se calcule dès lors à l'aide de la formule suivante :

$$S_E = \sqrt{\sum k_i p_i q_i\left(\frac{k_i}{k_i-1}\right)} \tag{4.59}$$

k_i = nombre d'items dans la strate *i*,

p_i = proportion d'items de la strate *i* réussis par le sujet,

$q_i = (1 - p_i)$

Comme le fait remarquer Felt (1984), cette formule risque malheureusement de conduire à des estimations très imprécises car les tests comprennent habituellement un grand nombre de strates comportant chacune un nombre relativement petit d'items. Lorsque certaines strates ne contiennent que deux ou trois items, Felt & al. (1985) conseillent d'ailleurs de ne pas utiliser cette formule pour estimer l'erreur type de mesure des différents scores d'un test.

7. L'étude de la généralisabilité

Les situations décrites jusqu'à maintenant ont illustré des cas relativement simples de calcul de la fiabilité dans le modèle classique : fiabilité de l'instrument de mesure en fonction du temps, de l'échantillonnage des items, etc. Il arrive, cependant, que les conditions d'observation et de mesure soient beaucoup plus complexes. Le problème se pose alors d'étudier la fiabilité à l'intérieur d'une famille de situations ou si l'on préfère d'un *univers de généralisabilité*. Dans un tel contexte, la notion de *score vrai* cède la place à la notion de *score univers*, score attendu de l'individu dans un ensemble de conditions d'observation et de mesure.

Prenons un exemple pour illustrer tout l'intérêt de l'étude de généralisabilité. Nous savons que la correction de compositions écrites présente un défi majeur aux enseignant(e)s. Il n'est pas facile d'obtenir des résultats fiables lors de la correction car plusieurs facteurs peuvent affecter la notation de l'élève. Il y a d'abord le sujet imposé de la composition écrite. Ensuite, il y a le degré de sévérité et de constance de chaque correcteur. Enfin, si chaque correcteur utilise une grille d'appréciation, la clarté et la facilité d'utilisation de la grille peuvent également influencer le travail de correction et de là, le score de l'élève. Comment traiter une telle situation avec les outils que nous avons vus dans ce chapitre, en particulier avec les coefficients de corrélation ?

D'abord, nous pourrions calculer plusieurs résultats pour chaque élève. Par exemple, chaque élève pourrait obtenir un score sur chaque thème imposé, pour chaque correcteur ou pour chaque grille de correction. Afin de déterminer la fiabilité inter-correcteurs, nous pourrions calculer les corrélations deux à deux entre les résultats accordés par chaque correcteur à chacun des thèmes. S'il ne devait y avoir que deux thèmes et trois correcteurs, nous devrions alors calculer six corrélations : les corrélations entre les correcteurs 1 et 2, 1 et 3, 2 et 3 pour le thème 1, et de même pour le thème 2. Si les corrélations entre les correcteurs devaient varier pour les résultats obtenus par les élèves aux deux thèmes, nous pourrions affirmer que la fiabilité inter-correcteurs est affectée par la nature du thème imposé. La nature du thème imposé serait considérée comme une source d'erreur de mesure.

Bien entendu, nous pourrions simplifier ce problème en ne calculant la fiabilité des résultats que pour les moyennes de chaque élève aux deux compositions écrites. Ceci pourrait améliorer la fiabilité, même s'il nous serait difficile d'estimer dans quelle mesure. Le principal bénéfice de cette procédure serait, par contre, de simplifier le calcul de la fiabilité inter-correcteurs. En calculant des scores moyens pour les thèmes, il ne nous resterait que trois coefficients de corrélation à calculer entre les correcteurs 1 et 2, 1 et 3 et 2 et 3. Mais que pourrions-nous dire maintenant de l'effet de la grille d'appréciation utilisée par les correcteurs ?

Là encore, la procédure à suivre risquerait d'être longue. En limitant à deux le nombre de grilles, nous voudrions sans doute nous assurer de la fiabilité des résultats obtenus en calculant, pour chaque correcteur, une corrélation entre les résultats accordés sur chacun des deux thèmes par les deux grilles. En effet, il faudrait calculer, pour chacun des trois correcteurs, 6 coefficients de corrélation : les corrélations entre les résultats aux grilles 1 et 2 pour le thème 1 et de même pour le thème 2. Que faire si les corrélations entre les résultats aux grilles 1 et 2 devaient différer sensiblement pour le

thème 1 et le thème 2 ? Ceci indiquerait que l'une des grilles d'appréciation donne lieu à des résultats plus fiables lorsque les compositions des élèves portent sur un thème particulier. Comment réduire cette source d'erreur de mesure et comment savoir quelle part de cette erreur dépend des correcteurs eux-mêmes ?

7.1 NOTION DE SCORE UNIVERS

Cronbach, Gleser et Rajaratnam (1963) ont élaboré la théorie de la généralisabilité dans le but de réunir en un seul concept les différentes définitions de la fiabilité. En utilisant les principes de l'analyse de variance, Cronbach et al. proposent de quantifier l'importance de chaque source de variation d'une situation de mesure. Le score vrai devient l'espérance mathématique de toutes les observations possibles et l'erreur est le résultat d'une fluctuation dans l'échantillonnage de certains niveaux des facettes considérées (évaluateurs, moments, formes d'items, etc.).

La généralisabilité est donc un concept plus englobant que celui de fiabilité. Il permet de décrire des situations de mesure plus complexes et plus près de la réalité. Cardinet et Tourneur (1985 ; p. 23) la définissent ainsi :

> « La généralisabilité est donc le degré auquel on peut généraliser d'un résultat obtenu dans des conditions particulières à la valeur théorique recherchée ».

Les sources d'erreur de mesure dans un dispositif complexe sont fort nombreuses. L'étude de la fiabilité de tels dispositifs doit tenir compte de toutes les facettes du plan d'observation et de leurs interactions. Pour y arriver, il faut calculer la variabilité des résultats en fonction de ces différentes *facettes*. C'est donc de la fiabilité du *score univers* dont il sera question, c'est-à-dire de la fiabilité du score dans l'univers des conditions décrites par les facettes du plan d'observation.

Cardinet et Tourneur (1985 ; p. 23) définissent ainsi le score univers :

> « Le score univers d'une personne p, donnée idéale, représente la moyenne des scores de la personne p, calculée sur toutes les observations admissibles. Or l'observateur utilise le score observé, ou une fonction du score observé, pour estimer la valeur du score univers. Il généralise ainsi de l'échantillon à la population ».

Il y a donc un parallèle important entre *fiabilité* et *généralisabilité*. Dans le modèle classique, la fiabilité se définit en terme de corrélation entre le score observé et le score vrai. Plus la corrélation entre les deux est élevée, plus la fiabilité est grande. Il en va de même avec la notion de généralisabilité. Elle traduit le degré de corrélation entre le score observé et le score univers de l'individu. Plus cette corrélation est élevée, plus le score observé de l'individu ressemble à celui qu'il aurait obtenu s'il avait été soumis à l'ensemble des conditions de l'univers de généralisation.

Nous ne connaissons pas le score univers directement, mais nous pouvons l'estimer. Dans l'exemple précédent, la moyenne des résultats de l'élève pour les deux thèmes, notés au moyen de deux grilles différentes par trois correcteurs constituerait le score observé de l'élève. Ce score observé constitue une bonne estimation du score univers de l'élève jusqu'à un certain point. Si le dispositif de mesure constitue un bon échantillon des thèmes, des correcteurs et des grilles de correction, alors le score observé sera représentatif de la population des conditions de mesure et sa généralisabilité sera élevée. Nous pourrions aussi affirmer que la généralisabilité du score dépend

de la corrélation qui existe entre le score univers (ou score vrai) et le score observé dans les mêmes conditions d'observation et de mesure.

Immédiatement, une conclusion s'impose : plus l'échantillon des conditions d'observation se rapproche de la population, plus la généralisabilité sera grande. Dans notre exemple, si nous augmentions le nombre de thèmes, de correcteurs et de grilles, l'échantillon serait plus important et la généralisabilité du score plus grande. Mais, comment s'assurer d'une bonne généralisabilité ? Toutes les facettes sont-elles aussi importantes les unes que les autres ? Comment développer un dispositif de mesure qui soit économique et efficace ? Pouvoir répondre à ces questions est la motivation première des études de la théorie de la généralisabilité.

7.2 ÉTUDES G ET D

L'étude de la généralisabilité permet de tenir compte de multiples sources d'erreur dans l'estimation de la fiabilité. Comme nous venons de le voir, dès que nous sommes intéressés à généraliser à un grand nombre de conditions d'observation, le recours aux coefficients de corrélation pour rendre compte de la variabilité des résultats devient rapidement impraticable. Pour tenir compte de l'ensemble des variations qui se produisent dans un plan d'observation et des interactions possibles entre les facettes de ce plan, l'étude de la généralisabilité se fonde sur l'analyse de la variance (voir chapitre 2). Tout comme l'analyse de la variance permet un test d'hypothèse sur plus de deux groupes à la fois, l'étude de généralisabilité permet d'estimer l'importance des variations introduites par plus d'une variable ou facette. L'étude de la généralisabilité est donc au calcul de la fiabilité, ce que l'analyse de variance est au test t. Pas étonnant alors de retrouver l'analyse de variance à la base des méthodes de calcul de la généralisabilité.

Tout d'abord, il y a lieu de faire une distinction importante entre les deux finalités de l'étude de généralisabilité : *étude G* et *étude D*. Cette distinction est rendue nécessaire du fait que l'étude de la généralisabilité permet un plus grand contrôle sur les sources d'erreur de notre dispositif d'observation. Il est donc possible de faire beaucoup plus que de calculer l'indice de fiabilité d'un score univers (ou *coefficient de généralisabilité*). Le chercheur peut aussi estimer dans quelles conditions son dispositif d'observation présentera des conditions optimales.

Le parallèle entre études G et D et la théorie classique des tests est difficile à établir, mais il est possible. Lors du calcul de la fiabilité d'équivalence, le chercheur peut estimer combien d'items parallèles aux items de son test de départ il doit ajouter pour obtenir une fiabilité acceptable. Nous avons vu que la formule de Spearman-Brown (équation 4.47) nous permettait de faire ce calcul. Cette estimation de la nouvelle fiabilité du test obtenue à partir des résultats aux items du test de départ correspond à une étude D. Le calcul de la fiabilité du test de départ au moyen de la corrélation entre les deux formes parallèles du test correspond à l'étude G.

L'étude de la généralisabilité serait d'un intérêt pratique limité si elle se limitait à traduire au moyen d'un coefficient unique le degré de fiabilité du score univers dans un plan complexe d'observation. À quoi bon connaître l'importance des différentes sources de variation et d'erreur de mesure — ce qui est le propre de l'étude G — si

l'on ne prend pas le soin de les contrôler — ce qui est le propre de l'étude D — afin de s'assurer d'une meilleure fiabilité ou généralisabilité ?

Les limites du modèle classique du score vrai proviennent de la difficulté à préciser les sources de variation qu'il faut contrôler pour diminuer l'erreur de mesure. Dans l'exemple de départ où nous avions, en plus des élèves, trois sources importantes de variation des résultats (les correcteurs, le thème de la composition écrite, la grille d'appréciation), seule une étude de la généralisabilité permet de déterminer la part que jouent chacune de ces trois sources de variation et chacune de leurs interactions dans la variance d'échantillonnage globale.

7.3 LES QUATRE ÉTAPES D'UNE ÉTUDE DE GÉNÉRALISABILITÉ

Cardinet et Tourneur (1985) ont étendu la théorie de la généralisabilité initiale telle que formulée par Cronbach, Gleser, Nanda et Rajaratnam (1972). En effet, pour Cronbach et al, la facette « sujets » constitue le seul objet de mesure utile. Or, en psychologie et en éducation, le chercheur n'est pas uniquement intéressé par la stabilité des scores des sujets. Il s'intéresse aussi à la stabilité des effets d'autres objets de mesure tels que les items. Il peut s'agir d'estimer la stabilité des effets de différentes tâches ou de différentes modalités de présentation des items introduits dans un plan d'observation. Dans de telles conditions, ce ne sont plus les sujets que l'analyse de généralisabilité cherchera à différencier, mais bien les tâches et les contenus en tant qu'objets d'observation.

Cardinet et Tourneur (1985) ont donc défini une série de procédures de calcul applicables à tous les types de plans expérimentaux et qui permettent de tenir compte de tous les *projets de mesure*. En effet, selon ces auteurs (page 31) :

> L'erreur n'apparaît que par rapport à un projet de mesure. Elle suppose une intention particulière qui privilégie une ou plusieurs facettes comme conditions d'observation, c'est-à-dire comme sources d'erreurs... C'est (...) après le choix d'une direction privilégiée de mesure, que s'insère la théorie de la généralisabilité. Son rôle est de préciser l'importance de la variance due aux facettes privilégiées (variance de différenciation) par rapport à la variance due à l'échantillonnage des conditions d'observation (variance d'erreur).

La procédure proposée par Cardinet et Tourneur (1985) s'effectue en quatre étapes : les phases 1 et 2 se rapportent à l'analyse de variance ; la phase 3 se rapporte à l'étude G et la phase 4 à l'étude D. Voici une courte description de ces quatre étapes :

1. PLAN D'OBSERVATION : on procède au choix des facettes et du nombre de niveaux de chaque facette. On précise également les interrelations (*nichage, croisement*) entre ces facettes.
2. PLAN D'ESTIMATION : on détermine quelles facettes représentent un ensemble de niveaux fini ou infini et quelles facettes sont échantillonnées de façon *aléatoire* ou *exhaustive* (*effet fixe*).
3. PLAN DE MESURE : on identifie quelles facettes sont liées au projet de mesure (*facettes de différenciation*) et quelles facettes sont considérées comme sources d'erreur de mesure (*facettes d'instrumentation*). C'est au cours de cette phase que les composantes de variance calculées à la phase deux peuvent être attribuées à la variance vraie ou à la variance d'erreur, permettant ainsi le calcul du

coefficient de généralisabilité et le calcul de marges d'erreur applicables aux scores observés.

4. PLAN D'OPTIMISATION : cette phase consiste à modifier soit le plan d'observation, soit le plan d'estimation, soit le plan de mesure ou encore une combinaison des trois afin de maximiser la généralisabilité des observations. Le chercheur devra trouver alors un équilibre entre précision de la mesure et étendue de l'univers de généralisation. En effet, plus l'univers de généralisation est grand, plus il est difficile d'obtenir des mesures proches du score univers. Par contre, il y a peu d'intérêt pratique à utiliser des mesures précises lorsque l'univers de généralisation est trop étroit.

Dans notre exemple de départ, le plan d'observation est constitué de quatre facettes : les élèves, les thèmes des compositions écrites (2), les correcteurs (3) et les grilles de correction (2). Ces facettes sont totalement *croisées* si tous les élèves écrivent sur les deux thèmes et que chaque thème est corrigé par les trois correcteurs utilisant à chaque fois deux grilles de correction. Il serait possible d'agencer autrement les facettes de ce plan d'observation. Par exemple, il serait possible de *nicher* la facette « correcteur » dans la facette « thème » : deux correcteurs pourraient corriger le thème 1 au moyen des deux grilles pour chaque élève et deux autres correcteurs corrigeraient le thème 2 de la même façon. Nous dirions alors que la facette « correcteurs » est nichée dans la facette « thèmes », car les deux thèmes ne sont pas corrigées par tous les correcteurs. Une telle façon de procéder se justifie lorsque l'on souhaite attribuer la notation de chaque thème aux correcteurs les plus expérimentés.

Le plan d'estimation de notre exemple nous amène à définir le mode d'échantillonnage de nos facettes. Les élèves peuvent être considérés comme ayant été tirés au hasard d'une population infinie ou finie (si l'on en connaît la taille comme c'est le cas des élèves appartenant à un même district scolaire). En ce qui concerne les autres facettes, le plan d'estimation peut être plus délicat à établir. Les correcteurs peuvent aussi être considérés comme tirés d'une population finie ou infinie de correcteurs. Cette population serait considérée comme finie si l'on connaissait tous les enseignants susceptibles de corriger les épreuves. Les grilles de correction peuvent être considérées comme ayant été tirées d'une infinité de possibilités de grilles. Nous pouvons aussi considérer comme fixe cette facette et ne souhaiter généraliser les résultats des élèves qu'à deux grilles. Cette procédure serait adéquate si, d'année en année, les deux mêmes grilles étaient réutilisées. Quant aux deux thèmes, les mêmes choix s'imposent : voulons-nous généraliser les résultats des élèves à ces deux seuls thèmes ou à tous les thèmes ? Il peut être difficile de définir la population des thèmes : le programme d'études peut en prévoir un certain nombre. Dans ce cas, il serait possible de considérer les thèmes comme ayant été tirés d'une population finie, si notre but est de généraliser à l'ensemble des thèmes définis par le programme d'études. On pourrait justifier une telle procédure si d'une année à l'autre, deux nouveaux thèmes étaient tirés au hasard de l'ensemble des thèmes de composition écrite prévus au programme d'études.

Pour simplifier la situation, nous considérerons que tous les niveaux de facettes ont été tirés de populations infinies. Ceci aura pour effet de simplifier le calcul des composantes de variance. Dans le cas où les niveaux d'une ou plusieurs facettes devai-

ent être tirés d'une population finie ou encore représenter un échantillon exhaustif de tous les niveaux, le calcul des composantes de variance se ferait différemment.

Dans le plan de mesure, nous devons préciser la ou les facette(s) liées à notre projet de mesure. Si c'est le score de chaque élève en composition écrite qui nous intéresse, alors la facette « sujets » sera considérée comme facette de différenciation et les facettes « correcteurs », « thèmes » et « grilles de correction » comme des facettes d'instrumentation. Par contre, si c'est la fiabilité des correcteurs qui nous préoccupe, c'est la facette « correcteurs » qui deviendra facette de différenciation. La facette « sujets » sera alors considérée comme facette d'instrumentation avec les deux autres facettes. En effet, dans cette perspective, la fiabilité des résultats octroyés par les correcteurs dépend des variations que les sujets introduisent dans la qualité de leurs productions écrites.

Une fois le calcul des composantes de variance terminé (ou étude G), nous pouvons passer à une quatrième étape : le plan d'optimisation ou étude D. Cette dernière étape nous permettra d'entrevoir différents moyens d'améliorer notre dispositif de mesure.

7.4 REPRÉSENTATION GRAPHIQUE DES COMPOSANTES DE VARIANCE

La figure 7 présente les sources de variation du plan d'observation de l'exemple initial pour deux des trois facettes : les correcteurs (C) et les thèmes (T). Le *diagramme de Cronbach* est employé pour représenter graphiquement les sources de variation et leurs interactions. La facette de différenciation « sujets » (S) y est illustrée en gris, en plus des facettes d'instrumentation C et T (« correcteurs » et « thèmes ») en blanc. Elles sont entièrement croisées avec la facette S. Les aires d'intersection entre les ellipses représentent les interactions entre facettes.

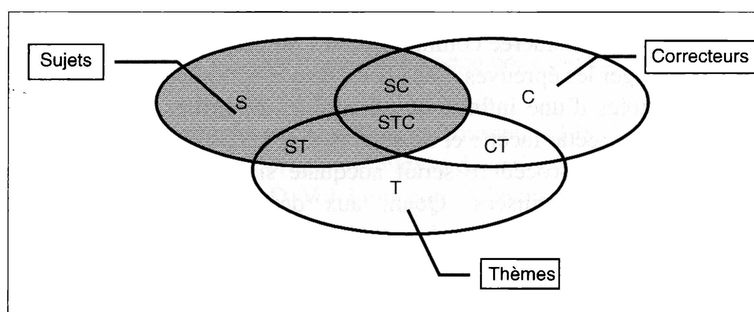


Figure 7 – Diagramme de Cronbach de trois facettes croisées

La figure 8 présente les sources de variation du plan d'observation lorsque les correcteurs sont nichés sous chacun des deux thèmes. Le nichage des facettes est représenté par l'inclusion d'une ellipse (*facette nichée*) dans une ellipse plus grande (*facette nichante*). La relation de nichage est indiquée par les deux points « : ». Ainsi, C :T signifie que la facette correcteurs est nichée dans la facette thèmes. Ce nouveau plan d'observation rend impossible l'identification d'une composante d'interaction SC

indépendante de la composante T. Du fait que la facette C est maintenant nichée dans T, la composante C ne peut plus être distinguée de la composante CT. De même, le nouveau plan d'observation rend impossible l'identification d'une composante d'interaction SC indépendante de la composante SCT, car les deux sont confondues dans la composante SC :T.

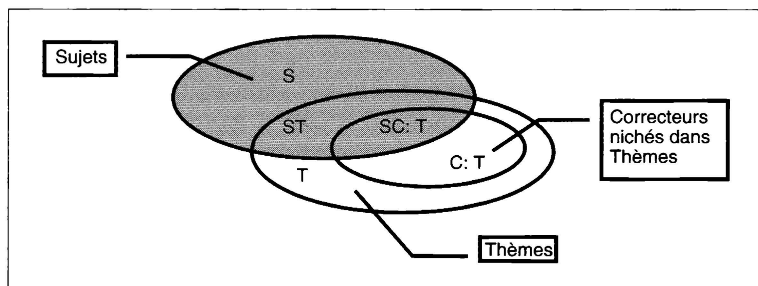


Figure 8 – Diagramme de Cronbach illustrant deux facettes nichées

La figure 9 nous montre de plus près les composantes de variance entrant dans la composition de la variance de la facette »sujets » pour le plan d'observation de l'exemple initial (figure 7). Nous retrouvons à l'intérieur de l'ellipse de la facette S (en gris) des composantes de variation partagées avec les facettes d'instrumentation ou *univers de généralisation.* En effet, les résultats des élèves ne dépendent pas que de leurs différences individuelles. Si les correcteurs ont été moins sévères envers certains élèves, cette interaction SC entrera comme composante de la variation entre les sujets. De même, si le thème 1 s'avère plus facile pour certains élèves, alors que le thème 2 est plus facile pour d'autres, cette nouvelle interaction ST s'ajoutera au sources de variation. Enfin, il est possible que selon le correcteur et l'élève, la composition écrite sous un thème soit notée plus ou moins sévèrement. Cette triple interaction STC s'ajoute à nouveau aux sources de variation entre les sujets. Toutes ces sources de variation s'accumulent comme composantes plus ou moins grandes de la variation entre les sujets et constituent autant de sources d'erreur qui masquent les différences réelles entre les sujets. Comment intervenir dans ces circonstances pour améliorer la généralisabilité.

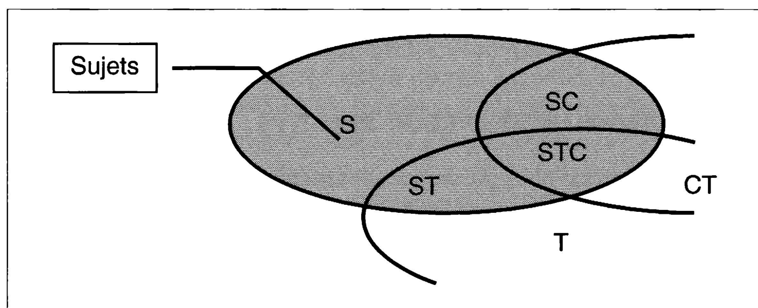


Figure 9 – Diagramme de Cronbach des composantes de variance de la facette S

Plusieurs possibilités se présentent. Devons-nous accroître le nombre de correcteurs ? Serait-il préférable de réduire le nombre de correcteurs mais d'accroître le nombre de thèmes des compositions écrites réalisées par chaque élève ? Deux grilles d'appréciation sont-elles nécessaires ? Voilà autant de points sur lesquels une décision doit être prise et où l'étude D est susceptible de rendre de précieux services. Pour ce faire, il nous faut connaître l'importance de ces sources de variations. C'est ce que permettra de réaliser l'étude G des composantes de variance.

Dans une situation idéale, la plus grande part de la variance entre les sujets dépendrait uniquement des sujets. Les interactions « correcteurs × sujets » et « thèmes × sujets », considérées comme des sources d'erreur, ne représenteraient qu'une petite proportion de la variance totale entre sujets. La mesure est au contraire insatisfaisante lorsqu'une grande proportion de la variation entre les sujets est imputable à ces interactions. Tant dans le modèle classique que dans l'étude de la généralisabilité, la fiabilité est calculée à partir de la proportion de la variance observée qui est due à la variance des scores vrais. Dans le contexte de la théorie de la généralisabilité, la variance due aux scores vrais est remplacée par ce qu'il est convenu d'appeler la *variance de différenciation* ou si l'on préfère *la variance attendue des scores univers.*

7.5 REPRÉSENTATION SYMBOLIQUE

L'indice de fiabilité tiré de l'étude de la généralisabilité se définit donc simplement comme la proportion de la variance des scores observés résultant de la variance de différenciation :

$$\rho^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_e^2} \qquad (4.60)$$

Dans l'équation précédente, σ_τ^2 représente la variance de différenciation ou variance attendue des scores univers et $(\sigma_\tau^2 + \sigma_e^2)$ représente la variance attendue des scores observés. Par définition, c'est la somme de la variance des scores univers et de la variance des erreurs d'échantillonnage. Le terme d'erreur σ_e^2 dépend de plusieurs facteurs. Intuitivement, il est facile de comprendre que plus l'univers de généralisation est grand, plus ce terme risque d'être élevé. Enfin, selon que nous sommes intéressés par la valeur absolue du score univers (comme en évaluation critériée, voir chapitre 3) ou par sa valeur relative (comme en évaluation normative, voir chapitre 3 et aussi le chapitre 7), la composante d'erreur sera différente.

7.6 ERREUR ABSOLUE ET ERREUR RELATIVE

Deux types d'erreur nous préoccupent particulièrement lorsqu'il s'agit de fiabilité de la mesure : l'erreur relative et l'erreur absolue. L'erreur relative se produit lorsque la position des résultats les uns par rapport aux autres se trouve changée. L'erreur absolue se produit lorsque la valeur absolue des résultats, telle que mesurée sur une échelle dont les échelons sont définis a priori, est changée. Dans un concours ou une évaluation de type sélection, l'erreur absolue n'a pas d'importance : il s'agit de ne

sélectionner que les meilleurs, quel que soit le score obtenu par chaque participant(e). Par contre, dans une épreuve de certification ou pour être admis dans une profession ou un programme d'études contingenté, la valeur absolue du résultat est également importante. Ce n'est pas la position relative du score par rapport aux autres scores qui nous préoccupe, mais c'est davantage la position de ce score par rapport à un seuil de réussite. Il ne serait pas approprié de permettre à quelqu'un de conduire un véhicule automobile sur la seule base qu'il s'est avéré le conducteur le moins mauvais parmi ceux qui se sont présentés. Pour obtenir un permis de conduire, le conducteur en question doit démontrer une performance minimale. L'erreur absolue fait intervenir à la fois les composantes d'erreur relative et d'erreur absolue.

Cette distinction entre erreur relative et erreur absolue est essentielle en psychologie et en éducation. Dans tout système de mesure où des seuils critiques sont utilisés pour déterminer si un stade a été atteint, une étape de développement franchie, un seuil de maîtrise réussi, l'erreur absolue de mesure joue un rôle important. En psychologie différentielle, c'est l'erreur sur les positions relatives qui est la plus pertinente. Par exemple, lorsque les tests d'aptitude sont utilisés à des fins de sélection, l'erreur relative prime. Le directeur d'école qui souhaite créer une classe regroupant les meilleurs élèves n'est pas préoccupé par l'erreur absolue. Il lui importe de sélectionner les 25 meilleurs candidats pour cette classe quelle que soit la valeur absolue de leurs résultats. Si pour créer une telle classe, chaque élève devait avoir un QI de 120 et plus, il se pourrait qu'il ne trouve dans son école que peu d'élèves de ce niveau et ne puisse créer la classe projetée. Il lui serait alors impossible de créer une classe à voie enrichie avec le seuil de réussite fixé.

7.7 EXEMPLE

Nous ne présenterons pas ici les détails des procédures de calcul intervenant dans l'étude de généralisabilité. Il faut pour cela une connaissance approfondie de l'analyse de variance et de l'estimation statistique qui dépassent les prérequis de cet ouvrage. Il est possible, par contre, de saisir l'utilité de l'étude de généralisabilité à travers une simulation qui illustre sa capacité à apporter des solutions satisfaisantes à bon nombre de problèmes courants impliquant la fiabilité de la mesure.

Cette simulation aura comme principal avantage de nous permettre de connaître *a priori* les effets introduits par les principales facettes impliquées dans la variation des résultats. Nous serons donc à même de constater comment l'étude de la généralisabilité permet de retrouver les principaux effets et leurs interactions introduits dans les données de départ et comment ceux-ci affectent l'estimation de la fiabilité.

La situation que nous chercherons à décrire est celle de la fiabilité des notes accordées par des juges à une série de plongeons aux figures imposées. Cette situation est représentée graphiquement par le diagramme de Cronbach de la figure 10. Comme on peut le constater, trois principales sources d'erreur relative sont en jeu : la possibilité que les juges notent différemment un même sujet (SJ), la possibilité qu'un même sujet éprouve des difficultés particulières pour un plongeon plutôt qu'un autre (SP) et enfin, la possibilité que les juges notent différemment des plongeons en fonction de chaque sujet (SJP). Si notre objectif se limite à classer les plongeurs et à décerner trois médailles (or, argent et bronze), ces sources d'erreur sont les seules qui devraient nous

préoccuper car elles affectent la position relative d'un plongeur par rapport à un autre. Il nous importe peu de savoir si le médaillé d'or se mérite 7,4 plutôt que 6,9. L'essentiel est que son score soit le plus élevé, quel que soit le juge qui l'ait noté ou le plongeon qu'il ait exécuté.

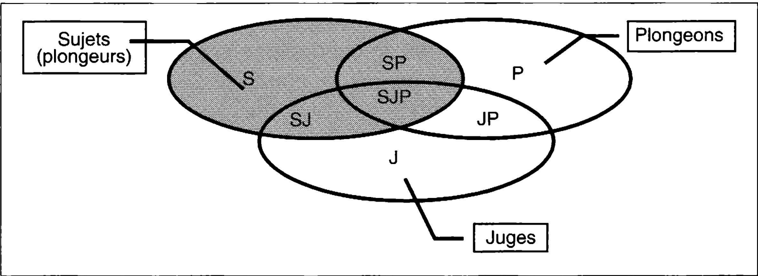


Figure 10 – Diagramme de Cronbach du plan d'observation de la simulation

Tableau 9 – Données de départ de la simulation

Sujets	Score univers	J1 P1	J1 P2	J1 P3	J2 P1	J2 P2	J2 P3	J3 P1	J3 P2	J3 P3	Score observé
1	3,01	3,01	3,01	3,01	3,01	3,01	3,01	3,01	3,01	3,01	3,01
2	6,30	6,30	6,30	6,30	6,30	6,30	6,30	6,30	6,30	6,30	6,30
3	3,91	3,91	3,91	3,91	3,91	3,91	3,91	3,91	3,91	3,91	3,91
4	5,49	5,49	5,49	5,49	5,49	5,49	5,49	5,49	5,49	5,49	5,49
5	4,51	4,51	4,51	4,51	4,51	4,51	4,51	4,51	4,51	4,51	4,51
6	5,77	5,77	5,77	5,77	5,77	5,77	5,77	5,77	5,77	5,77	5,77
7	4,76	4,76	4,76	4,76	4,76	4,76	4,76	4,76	4,76	4,76	4,76
8	4,21	4,21	4,21	4,21	4,21	4,21	4,21	4,21	4,21	4,21	4,21
9	5,06	5,06	5,06	5,06	5,06	5,06	5,06	5,06	5,06	5,06	5,06
10	5,68	5,68	5,68	5,68	5,68	5,68	5,68	5,68	5,68	5,68	5,68
11	6,21	6,21	6,21	6,21	6,21	6,21	6,21	6,21	6,21	6,21	6,21
12	6,18	6,18	6,18	6,18	6,18	6,18	6,18	6,18	6,18	6,18	6,18
Moyenne	5,09	5,09	5,09	5,09	5,09	5,09	5,09	5,09	5,09	5,09	5,09
Écart-type	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99

Par contre, si la valeur absolue du score est importante, il faudrait tenir compte de sources d'erreur additionnelles. Pour être admis comme sauveteur, il ne suffit pas d'être le meilleur de son groupe. Il faut aussi exécuter les plongeons avec un certain degré de maîtrise. La valeur absolue de la performance devient primordiale. En plus des sources d'erreur relative, nous devrions tenir compte de la sévérité des juges et de la difficulté des plongeons retenus. Si, par hasard, trois juges particulièrement sévères

étaient choisis, les résultats de tous les plongeurs seraient sous-estimés. Si, de la même manière, les exercices comptaient parmi les figures imposées les plus difficiles, là encore les plongeurs risqueraient de rater le seuil de maîtrise fixé.

Le tableau 9 présente la situation des plongeurs avant que nous introduisions les effets pour les facettes principales (juges et plongeons) et leurs interactions. Ces douze plongeurs ont été tirés au hasard d'une population où le score univers moyen vaut 5 et la variance des scores univers vaut 1. En l'absence d'écarts introduits par la sévérité des juges ou par la difficulté des plongeons, le score observé demeure identique au score univers pour chaque plongeur. Nous sommes dans une situation où, à l'évidence, ni les plongeons, ni les juges ne sont une source d'erreur aléatoire. Cette situation nous conduirait à une généralisabilité parfaite du résultat des plongeurs, puisque celui-ci demeurerait le même peu importe le juge ou le plongeon exécuté. Cette situation, quoiqu'idéale, n'est pas réaliste.

Le tableau 10 introduit des effets pour les juges et pour les plongeons. Le juge 2 est le moins sévère, car il alloue 1, 5 points de plus à tous les plongeurs. Les juges 2 et 3, plus sévères, accordent quant à eux un résultat inférieur de -0, 75 à chaque plongeur. Quant aux plongeons, le premier est celui pour lequel les athlètes se voient accorder le plus de points (ou le plus facile), suivi des plongeons 2 et 3.

Tableau 10 – Effets des juges et des plongeons ajoutées au données de départ
(J1 = -0,75 ; J2 = +1,5 ; J3 = -0,75)
(P1 = +0,5 ; P2 = +0,25 ; P3 = -0,75)

Sujets	Score univers	J1	J1	J1	J2	J2	J2	J3	J3	J3	Score observé
		P1	P2	P3	P1	P2	P3	P1	P2	P3	
1	3,01	2,76	2,51	1,51	5,01	4,76	3,76	2,76	2,51	1,51	3,01
2	6,30	6,05	5,80	4,80	8,30	8,05	7,05	6,05	5,80	4,80	6,30
3	3,91	3,66	3,41	2,41	5,91	5,66	4,66	3,66	3,41	2,41	3,91
4	5,49	5,24	4,99	3,99	7,49	7,24	6,24	5,24	4,99	3,99	5,49
5	4,51	4,26	4,01	3,01	6,51	6,26	5,26	4,26	4,01	3,01	4,51
6	5,77	5,52	5,27	4,27	7,77	7,52	6,52	5,52	5,27	4,27	5,77
7	4,76	4,51	4,26	3,26	6,76	6,51	5,51	4,51	4,26	3,26	4,76
8	4,21	3,96	3,71	2,71	6,21	5,96	4,96	3,96	3,71	2,71	4,21
9	5,06	4,81	4,56	3,56	7,06	6,81	5,81	4,81	4,56	3,56	5,06
10	5,68	5,43	5,18	4,18	7,68	7,43	6,43	5,43	5,18	4,18	5,68
11	6,21	5,96	5,71	4,71	8,21	7,96	6,96	5,96	5,71	4,71	6,21
12	6,18	5,93	5,68	4,68	8,18	7,93	6,93	5,93	5,68	4,68	6,18
Moyenne	5,09	4,84	4,59	3,59	7,09	6,84	5,84	4,84	4,59	3,59	5,09
Écart-type	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99

Ces écarts introduits par les juges et les plongeons tendent à surestimer ou à sous-estimer à chaque notation l'habileté des plongeurs. Il en résulte une note supérieure ou inférieure au score univers de plongeon de chaque athlète. Dans notre exemple, afin de simplifier l'interprétation, la somme des erreurs d'estimation s'annule lorsque l'on prend en considération tous les juges et tous les plongeons. C'est pourquoi, même si les résultats individuels ont changé, leur moyenne par sujets demeure constante.

Il existe tout de même une erreur absolue sur chaque note. En effet, selon que l'on considère un juge plutôt qu'un autre, ou encore un plongeon plutôt qu'un autre, la note varie. Cette erreur absolue serait importante si le but de cet exercice était de déterminer ceux et celles qui ont atteint un seuil de performance qui les rend admissibles au métier de sauveteur. Si un seuil de 8 est exigé en plongeon, plusieurs plongeurs se verraient acceptés par certains juges pour certains plongeons, alors qu'ils auraient dû être refusés. Il s'agirait de *faux positifs* (voir chapitre 5) : ces plongeurs sont acceptés sur base de leur score observé, alors qu'ils devraient être refusés, étant donné leur score univers (ou score vrai).

Ce type d'erreur absolue n'a toutefois rien à voir avec le classement relatif des plongeurs. S'il s'agit d'une compétition devant déterminer les trois meilleurs, la sévérité des juges ou la facilité des plongeons n'ont aucun effet sur la position relative de chaque plongeur dans le classement. Si l'on additionne les points mérités par chaque plongeur, on observe que le plongeur 2 est toujours celui qui se mérite la moyenne la plus élevée. Les effets ajoutés ayant joué pour tous, le classement n'est pas affecté. Donc, la généralisabilité relative des résultats du tableau 10 demeure parfaite.

Une autre source d'erreur absolue pourrait se présenter si les juges, en plus d'être plus ou moins sévères entre eux, différaient quant aux résultats qu'ils accordent à chaque plongeon. Dans la situation précédente, le juge 2 accordait 1,5 de plus à chaque plongeur et ce, peu importe le plongeon. Il en allait de même pour les autres juges. Bref, tous les juges étaient constants dans leur degré d'indulgence ou de sévérité, peu importe le plongeon.

La matrice du tableau 11 nous indique les effets d'interaction entre les trois juges et les trois plongeons. Le juge 1, par exemple, accorde 0,5 point de plus au plongeon 3, alors que le juge 3 enlève -0,5 points au plongeon 1. Une tel comportement des juges 1 et 3 pourrait s'expliquer par le fait que ces deux juges évaluent différemment la complexité du plongeon. Le juge 3, considérant le plongeon 1 plus facile que les deux autres, est plus sévère pour ce plongeon. Le juge 1, considérant le plongeon 3 comme plus difficile, est plus indulgent.

Tableau 11 – Interaction juges X plongeons

	J1	J2	J3
P1	0	-0,25	-0,5
P2	0	0,25	0
P3	0,5	0	0

Ces écarts dus à l'interaction « juges x plongeons » ont été ajoutés aux résultats du tableau 10 pour donner les résultats du tableau 12. Puisque la somme de ces interactions est nulle et que chaque plongeur a été affecté également par l'effet de ces interactions, ni la valeur absolue de leur score individuel, ni le classement n'ont été affectés. Le plongeur 2 demeure toujours le champion. La seule erreur due à cette interaction est une erreur absolue dans l'estimation du score univers pour un juge et un plongeon particulier. L'erreur relative demeure nulle.

Tableau 12 – Ajout de l'interaction juges X plongeons aux données

Sujets	Score	J1	J1	J1	J2	J2	J2	J3	J3	J3	Score
	univers	P1	P2	P3	P1	P2	P3	P1	P2	P3	observé
1	3,01	2,76	2,51	2,01	4,76	5,01	3,76	2,26	2,51	1,51	3,01
2	6,30	6,05	5,80	5,30	8,05	8,30	7,05	5,55	5,80	4,80	6,30
3	3,91	3,66	3,41	2,91	5,66	5,91	4,66	3,16	3,41	2,41	3,91
4	5,49	5,24	4,99	4,49	7,24	7,49	6,24	4,74	4,99	3,99	5,49
5	4,51	4,26	4,01	3,51	6,26	6,51	5,26	3,76	4,01	3,01	4,51
6	5,77	5,52	5,27	4,77	7,52	7,77	6,52	5,02	5,27	4,27	5,77
7	4,76	4,51	4,26	3,76	6,51	6,76	5,51	4,01	4,26	3,26	4,76
8	4,21	3,96	3,71	3,21	5,96	6,21	4,96	3,46	3,71	2,71	4,21
9	5,06	4,81	4,56	4,06	6,81	7,06	5,81	4,31	4,56	3,56	5,06
10	5,68	5,43	5,18	4,68	7,43	7,68	6,43	4,93	5,18	4,18	5,68
11	6,21	5,96	5,71	5,21	7,96	8,21	6,96	5,46	5,71	4,71	6,21
12	6,18	5,93	5,68	5,18	7,93	8,18	6,93	5,43	5,68	4,68	6,18
Moyenne	5,09	4,84	4,59	4,09	6,84	7,09	5,84	4,34	4,59	3,59	5,09
Écart-type	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99

Les résultats des tableaux 10 et 12 nous présentent des valeurs constantes en termes de classement. Celui-ci est demeuré le même parce que sur l'ensemble des trois plongeons, les trois juges ont toujours eu la même « attitude » envers chaque plongeur. Si chaque juge devait accorder plus de points à un sujet en particulier à cause de critères subjectifs ou d'une interprétation personnelle des critères d'évaluation, il y aurait une interaction entre les juges et les sujets qui pourrait ressembler à ce que décrit le tableau 13.

On note dans le tableau 13 qu'en ce qui concerne les sujets 2, 5, 7 et 11, il n'y a eu aucune interaction. Par contre, le sujet 8 se voit accorder 1 point de moins par les juges 1 et 3 et 1 point de plus par le juge 2. Il en va de même des autres plongeurs, même si la grandeur des effets d'interaction peut varier. Ces effets d'interaction signifient qu'un juge a accordé plus de points ou moins de points à un plongeur particulier.

Le juge 2, celui qui accorde le plus de points à tous les plongeurs peu importe le plongeon, a donné un point de plus au sujet 8 et un point de moins au sujet 12. Le classement des plongeurs est affecté par de telles interactions. C'est là une source importante d'erreur relative. Les juges peuvent différer entre eux dans leur notation des plongeurs et se laisser influencer par des critères non objectifs.

Tableau 13 – Interaction juges X sujets

Sujets	J1	J2	J3
1	0	0	0,25
2	0	0	0
3	0,5	0,5	-0,25
4	0	0	-0,75
5	0	0	0
6	-0,5	-0,25	0,5
7	0	0	0
8	-1	1	-1
9	0	0	0,25
10	0	0	0,25
11	0	0	0
12	1	-1	0,5

Le tableau 14 présente les nouveaux résultats, une fois ajoutée l'interaction entre juges et sujets. Si l'on compare les scores observés, on constate que, pour la première fois, le classement des plongeurs a été affecté par ces effets d'interaction. En effet, le champion n'est plus le plongeur #2 (6,30), mais bien le plongeur #12 (6,35). Les scores individuels de chaque plongeur ont été affectés par ces interactions « juges x sujets », même si la somme de ces interactions, égale à 0, ne change pas la moyenne du groupe des 12 plongeurs. Si le classement des plongeurs est primordial, nous voudrions certainement réduire au minimum l'importance de ces erreurs relatives dans la variation des scores observés des sujets.

L'interaction « juges x sujets » n'est pas la seule source d'erreur relative qui puisse affecter le classement des sujets. Jusqu'ici, nous avons admis que la valeur relative des résultats obtenus à chaque plongeon était identique pour chaque plongeur. Un tel postulat serait admissible si, par exemple, le plongeon 1 était le plus facile et qu'il en allait de même pour tous les sujets. Mais, ce postulat se vérifie mal dans la réalité. Si un plongeon peut être le plus facile pour une majorité de sujets, il est possible que la difficulté relative de chaque plongeon varie d'un sujet à l'autre. C'est ce que tente d'illustrer la matrice d'interaction « sujets x plongeons » du tableau 15.

Tableau 14 – Ajout de l'interaction juges X sujets aux données

Sujets	Score univers	J1	J1	J1	J2	J2	J2	J3	J3	J3	Score observé
		P1	P2	P3	P1	P2	P3	P1	P2	P3	
1	3,01	2,76	2,51	2,01	4,76	5,01	3,76	2,51	2,76	1,76	3,10
2	6,30	6,05	5,80	5,30	8,05	8,30	7,05	5,55	5,80	4,80	6,30
3	3,91	4,16	3,91	3,41	6,16	6,41	5,16	2,91	3,16	2,16	4,16
4	5,49	5,24	4,99	4,49	7,24	7,49	6,24	3,99	4,24	3,24	5,24
5	4,51	4,26	4,01	3,51	6,26	6,51	5,26	3,76	4,01	3,01	4,51
6	5,77	5,02	4,77	4,27	7,27	7,52	6,27	5,52	5,77	4,77	6,68
7	4,76	4,51	4,26	3,76	6,51	6,76	5,51	4,01	4,26	3,26	4,76
8	4,21	2,96	2,71	2,21	6,96	7,21	5,96	2,46	2,71	1,71	3,88
9	5,06	4,81	4,56	4,06	6,81	7,06	5,81	4,56	4,81	3,81	5,14
10	5,68	5,43	5,18	4,68	7,43	7,68	6,43	5,18	5,43	4,43	5,76
11	6,21	5,96	5,71	5,21	7,96	8,21	6,96	5,46	5,71	4,71	6,21
12	6,18	6,93	6,68	6,18	6,93	7,18	5,93	5,93	6,18	5,18	6,35
Moyenne	5,09	4,84	4,59	4,09	6,86	7,11	5,86	4,32	4,57	3,57	5,09
Écart-type	0,99	1,17	1,17	1,17	0,85	0,85	0,85	1,19	1,19	1,19	1,00

Dans ce tableau, on constate que le plongeon 1 s'avère le plus difficile des trois pour le sujet 9. Par contre, pour le sujet 1, c'est le plus facile. Dans l'ensemble, peu de sujets semblent affectés par cette interaction. Pour huit des 12 plongeurs, la difficulté relative de chaque plongeon ne change pas. Cette interaction peut-elle être considérée comme négligeable pour l'ensemble des sujets ?

Tableau 15 – Interaction sujets X plongeons

Sujets	P1	P2	P3
1	0,5	0,25	-0,75
2	0	0	0
3	0	0	0
4	0	0	0
5	0,25	0,25	-0,25
6	0	0	0
7	0	0	0
8	0	0	0
9	-0,5	-0,25	0
10	0	0	0
11	0	0	0
12	0	0	0,5

Nous pourrions ajouter encore la triple interaction « juges × plongeons × sujets ». Nous postulerons que celle-ci est nulle pour toutes les combinaisons de facettes. Le tableau 16 présente les résultats des 12 plongeurs une fois la double interaction « sujets × plongeons » ajoutée aux résultats du tableau précédent. L'effet sur le classement est sensible. Le plongeur dont le score univers était le plus élevé se classe maintenant second. La médaille d'or lui échappe à cause d'erreurs relatives de mesure occasionnées par les différentes interactions. Quant au vainqueur, le plongeur 12, son score univers de départ le classait troisième : le bronze s'est transformé en or pour ce plongeur grâce à une série d'erreurs relatives de mesure favorables.

Tableau 16 – Ajout de l'interaction plongeons X sujets aux données

Sujets	Score univers	J1	J1	J1	J2	J2	J2	J3	J3	J3	Score observé	Écart type
		P1	P2	P3	P1	P2	P3	P1	P2	P3		
1	3,01	3,26	2,76	1,26	5,26	5,26	3,01	3,01	3,01	1,01	3,10	1,38
2	6,30	6,05	5,80	5,30	8,05	8,30	7,05	5,55	5,80	4,80	6,30	1,15
3	3,91	4,16	3,91	3,41	6,16	6,41	5,16	2,91	3,16	2,16	4,16	1,38
4	5,49	5,24	4,99	4,49	7,24	7,49	6,24	3,99	4,24	3,24	5,24	1,38
5	4,51	4,51	4,26	3,26	6,51	6,76	5,01	4,01	4,26	2,76	4,59	1,26
6	5,77	5,02	4,77	4,27	7,27	7,52	6,27	5,52	5,77	4,77	5,68	1,07
7	4,76	4,51	4,26	3,76	6,51	6,76	5,51	4,01	4,26	3,26	4,76	1,15
8	4,21	2,96	2,71	2,21	6,96	7,21	5,96	2,46	2,71	1,71	3,88	2,05
9	5,06	4,31	4,31	4,06	6,31	6,81	5,81	4,06	4,56	3,51	4,89	1,05
10	5,68	5,43	5,18	4,68	7,43	7,68	6,43	5,18	5,43	4,43	5,76	1,09
11	6,21	5,96	5,71	5,21	7,96	8,21	6,96	5,46	5,71	4,71	6,21	1,15
12	6,18	6,93	6,68	6,68	6,93	7,18	6,43	5,93	6,18	5,68	6,52	0,47
Moyenne	5,09	4,86	4,61	4,05	6,88	7,13	5,82	4,34	4,59	3,53	5,09	1,22
Écart-type	0,99	1,10	1,13	1,38	0,76	0,79	1,04	1,12	1,15	1,36	1,01	0,34

7.8 ANALYSE DE VARIANCE ET ÉTUDE DE GÉNÉRALISABILITÉ

C'est à partir de l'analyse de variance que l'étude de la généralisabilité permet de déterminer les contributions relatives de chacune des facettes d'un dispositif de mesure, soit à la variance des scores univers *(variance de différenciation),* soit à la variance d'erreur relative ou absolue *(variance d'instrumentation).* Le calcul des différentes composantes de variance associées à un plan de mesure requiert une excellente connaissance de l'analyse de variance et des lois de l'estimation statistique. Pour plus de renseignements à ce sujet, le lecteur pourra consulter le livre de Cardinet et Tourneur (1985) qui précise toutes les étapes de ces calculs.

Une fois les calculs de composantes de variance effectués, l'étude de généralisabilité peut se poursuivre. Les résultats peuvent ressembler à ce que nous retrouvons au tableau 17 pour les données de la simulation présentées dans le tableau 16. On y trouve les résultats habituels de l'analyse de variance (sources de variance, degrés de liberté, carrés moyens). Dans les deux dernières colonnes, on y a ajouté des renseignements propres à l'étude de la généralisabilité : le calcul des composantes de variance exprimées en valeurs absolues et en pourcentages.

Tableau 17 – Analyse de variance et calcul des composantes de variance

Source de variation	Sommes des carrés	Degré de liberté	Carré moyen	Composantes de variance	%
S	110, 65	11	10, 059	1, 00341	30
J	127, 16	2	63, 578	1, 72855	51
SJ	18, 72	22	0, 851	0,28360	8
P	21, 22	2	10, 610	0, 27590	8
SP	3, 91	22	0, 178	0, 05917	2
JP	2, 00	4	0, 500	0, 04165	1
SJP	0, 00	44	0, 000	0, 00001	0

Les composantes de variance nous fournissent de précieuses informations en elles-mêmes. Elles nous indiquent quelles facettes sont responsables de la plus grande partie de la variance. En principe, nous devrions y retrouver les effets que nous avons introduits dans notre simulation. D'après le tableau 17, les composantes les plus importantes sont celles liées à la facette sujets (30%) et à celle des juges (51%). La composante de variance de la facette J est de beaucoup supérieure à celle de la facette des plongeons P. Les effets simples introduits pour la facette J sont de -0,75, +1,5 et -0,75 (une étendue de 2,25). Pour la facette P, il sont de +0,5, +0,25 et -0,75 (une étendue de 1,25). Que les résultats accordés par les juges constituent une composante de variance plus importante des résultats que le type de plongeon exécuté est donc conforme à notre modèle de simulation.

Parmi les composantes d'interaction les plus importantes, seule l'interaction SJ vaut la peine de s'y attarder. Elle représente 8% de la variance totale. Elle indique que les juges ne sont pas constants entre eux dans leur classement d'un même sujet. Pour un juge, un plongeur pourrait se mériter le meilleur score, alors que pour un autre juge, ce même plongeur pourrait se classer très différemment. Cette composante de variance est la seule source d'erreur relative vraiment importante. La composante de variance associée à l'interaction SP est bien moindre (étendue de -0,75 à +0,5) que celle due à l'interaction SJ (étendue de -1 à +1). Encore une fois, les résultats de l'analyse des composantes de variance est fidèle à notre modèle.

Les autres composantes de variance sont négligeables. La composante de variance associée à l'interaction juges x plongeons ne compte que pour 1,89% de la

variance. Les composantes associées à l'interaction SP et à la triple interaction SJP comptent pour 2% et 0%. Dans le cas de la triple interaction, le résultat de 0% n'est pas surprenant étant donné que nous n'avons pas introduit de tels effets de triple interaction dans notre modèle.

En résumé, nous retrouvons dans l'étude des composantes de variance les effets que nous avons introduits au départ. Les plus importants sont ceux liés à la facette « juges », à la facette « sujets » et à l'interaction « juges x sujets ». Il faut maintenant tenir compte des contributions respectives de ces facettes à la variance vraie (de différenciation) et à la variance d'erreur. C'est ici que débute véritablement l'étude de généralisabilité.

7.9 ÉTUDE G

Le tableau 18 regroupe les composantes de variance calculées en fonction de notre projet de mesure et de la nature de l'erreur (relative ou absolue) que nous souhaitons contrôler. Pour faciliter l'illustration de ces deux composantes de la variance dans le tableau 18, nous avons inscrit la variance de différenciation dans l'espace blanc et la variance d'erreur dans l'espace gris.

Comme notre projet de mesure consiste à différencier les plongeurs, la variance de différenciation sera constituée de la composante de variance des sujets. La variance d'instrumentation, lorsqu'il ne s'agit que de tenir compte de l'erreur relative de mesure, comprend toutes les composantes d'interaction impliquant la facette sujets avec les autres facettes : SJ, SP, SJP. Lorsqu'il s'agit d'erreur absolue, s'ajoutent aux composantes d'erreur relative précédentes, toutes les composantes de variance aléatoire des autres facettes d'instrumentation : J, P et leur interaction JP. Pour représenter le plan de mesure choisi, nous avons recours à la notation suivante : *(D/I)*. Dans ce système de notation, *D* représente la ou les facette(s) de différenciation (à gauche de la barre oblique) et *I* représente la ou les facette(s) d'instrumentation (à droite de la barre oblique). Cette notation ne tient compte que des facettes et non de leurs interactions ou nichages. Dans le cas de notre exemple, nous écrirons : (S/JP)

Tableau 18 – Analyse de généralisabilité pour le plan de mesure de départ (S/JP)

Source	Variance de différenciation	Source	Variance d'erreur relative	Variance d'erreur absolue
S	1, 00341			
		J		0, 57618
		SJ	0, 09453	0, 09453
		P		0, 09197
		SP	0, 01972	0, 01972
		JP		0, 00463
		SJP	0, 00000	0, 00000
Total (variance)	1, 00341		0, 11426	0, 78704
Écart type			0, 3380	0, 8872
Coefficient de généralisabilité			0, 898	0, 560

Le coefficient de généralisabilité se calcule selon l'équation (4.60). Si l'on substitue les valeurs des variances de différenciation et d'instrumentation du tableau 18 dans l'équation (4.60), nous retrouvons les valeurs des coefficients de généralisabilité relative et absolue.

$$\rho_\delta^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\delta^2} = \frac{1{,}00341}{1{,}00341 + 0{,}11426} = 0{,}898 \qquad (4.61)$$

$$\rho_\Delta^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\Delta^2} = \frac{1{,}00341}{1{,}00341 + 0{,}78704} = 0{,}560 \qquad (4.62)$$

Ce dernier est dénommé « *index of dependability* » et symbolisé par φ dans la littérature anglosaxonne. Ces résultats indiquent que la fiabilité des résultats est tout à fait acceptable lorsqu'il s'agit de classer les sujets. Un coefficient de généralisabilité relative de 0,898 indique une bonne fiabilité. Par contre, lorsqu'il s'agit d'utiliser la valeur absolue des scores, la fiabilité des résultats est moins satisfaisante (0,560). Si notre but premier était de situer les plongeurs par rapport à un seuil de réussite, nous aurions intérêt à diminuer l'erreur absolue des résultats.

La figure 11 illustre sous la forme d'un graphique circulaire la répartition des composantes de variance pour le plan de mesure de notre exemple. On y voit comment les composantes de variance du tableau 18 se répartissent entre la variance de différenciation et la variance d'instrumentation. On y constate que la principale source d'erreur relative est due à la composante d'interaction SJ. Pour améliorer le classement des sujets, nous devrions chercher à réduire cette composante. En ce qui concerne l'erreur absolue, la majeure partie vient de la composante des juges. La facette « juges » compte pour beaucoup dans la variance totale et en réduisant cette composante, nous pourrions améliorer la fiabilité absolue.

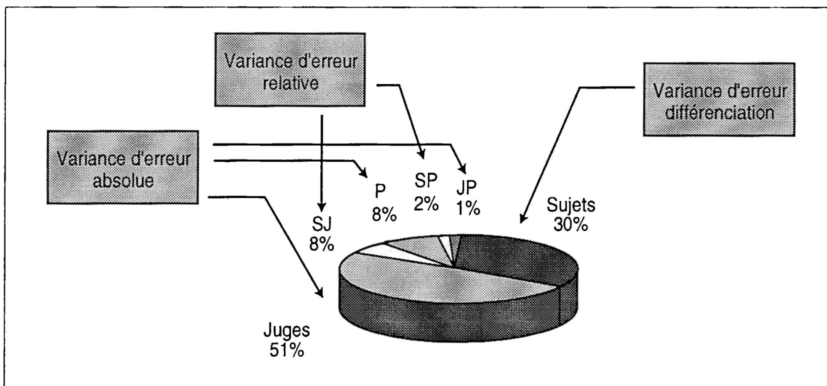


Figure 11 – Répartition des composantes de variance de l'exemple

On trouve au tableau 18 les valeurs des variances d'erreur relative et absolue, et à la ligne suivante les racines carrés de ces mêmes valeurs. Ces « erreurs types » représentent les écarts types de la distribution des fluctuations d'échantillonnage affectant les scores univers. Connaître ces écarts types nous permet d'établir autour de chaque

score observé un intervalle de confiance de ± 1,96 écarts types, marge à l'intérieur de laquelle on peut être à peu près certain que se situe le score univers, ou valeur vraie cherchée. Par exemple, l'étude G effectuée sur le dispositif d'évaluation utilisé pour les plongeurs nous permet d'affirmer que, pour un score observé de 6, le score vrai se situe entre 6 ± 1,96 × σ_δ, soit entre 5,34 et 6,66 (6 ± 1,96 × 0,3380) lorsqu'on ne s'intéresse qu'au classement relatif des plongeurs. L'intervalle de confiance serait presque trois fois plus étendu s'il s'agissait de déterminer la valeur absolue des performances de chaque plongeur.

Le tableau 19 récapitule toutes les étapes de notre simulation de données. On y retrouve les composantes de variance correspondant aux effets que nous avons introduits, de même que les valeurs calculées des coefficients de généralisabilité ρ^2_δ et ρ^2_Δ et des erreurs types (σ_δ et σ_Δ). Dans la situation de départ, 100% de la variance totale est due aux sujets. Les deux coefficients valent 1, car il n'y a ni erreur relative, ni erreur absolue : scores observés et scores univers correspondent parfaitement. En ajoutant des effets dus aux juges et aux plongeons, le classement des sujets demeure inchangé, car nous n'avons pas encore introduit d'interactions entre ces facettes et la facette sujets. C'est pourquoi la valeur de $\rho_\delta{}^2$ demeure inchangée à 1. La valeur de ρ^2_Δ passe par contre de 1 à 0,603, car les plongeons, mais surtout les juges, interviennent dans la valeur absolue des scores des plongeurs. Puisque les juges diffèrent entre eux dans leur notation, l'appréciation des plongeurs dépend en partie de ceux qui ont été choisis pour les évaluer. Trois juges ne constituent pas un très grand échantillon surtout lorsque les différences entre leurs appréciations sont si importantes.

L'introduction de l'interaction « juges × plongeons » ne change rien à la généralisabilité d'erreur relative. Celle-ci demeure parfaite, car cette interaction affecte tous les sujets de la même manière. En ce qui concerne l'erreur absolue, cette interaction devrait normalement contribuer à réduire encore plus la généralisabilité d'erreur absolue. Ce n'est pas ce qui s'est produit ici. L'interaction JP a réduit de façon importante la composante de variance P (de 14% à 8%), ce qui laisse un bilan positif en termes de composantes de variance d'erreur absolue. Nous sommes ici en présence d'un jeu contraire d'erreurs absolues.

La quatrième simulation introduit la composante d'interaction « sujets × juges ». Cette interaction change le classement des sujets et la généralisabilité d'erreur relative est maintenant de 0,913 : les juges manquent de constance entre eux dans leur appréciation des sujets. Cette erreur relative fait également partie de l'erreur absolue. C'est pourquoi le coefficient d'erreur absolue est également réduit à 0,566.

L'ajout, dans la cinquième simulation, d'une autre composante d'interaction « sujets × plongeons » ne changera que très peu les résultats de la quatrième simulation. Ces derniers résultats sont ceux que nous avons présentés dans les tableaux 17 et 18, ainsi que dans la figure 11. Cette interaction ne comptant que pour 1% de la variance totale, elle change peu de choses à la généralisabilité absolue ou relative. Les

sujets se classant de la même manière par rapport aux trois plongeons, cette interaction n'intervient donc que très peu dans la fiabilité des résultats.

Tableau 19 – Résultats à différentes étapes de la simulation des données

Simulation #	Effets introduits	Composantes de variance (%)							Généralisabilités et erreurs			
		S	J	SJ	P	SP	JP	SJP	ρ^2_δ	σ_δ	ρ^2_Δ	σ_Δ
1	NIL	100	0	0	0	0	0	0	1, 000	0, 000	1, 000	0, 003
2	J, P	34	53	0	14	0	0	0	1, 000	0, 000	0, 603	0, 842
3	JP	35	56	0	8	0	1	0	1, 000	0, 000	0, 622	0, 808
4	SJ	30	53	9	7	0	1	0	0, 913	0, 308	0, 566	0, 870
5	SP	30	51	8	8	2	1	0	0,898	0, 338	0, 560	0, 887

L'étude des composantes de variance pour le projet de mesure consistant à différencier les sujets nous a fourni quelques pistes quant aux meilleurs moyens d'améliorer la généralisabilité des résultats de notre dispositif de mesure. Les améliorations à apporter devront contribuer à réduire l'erreur absolue de mesure due à la composante « juges » et à la composante d'interaction « sujets × juges ». Si seule l'erreur relative nous importe, alors il suffira de réduire la composante d'interaction « sujets × juges » seulement. C'est ce que nous verrons dans l'étude D ou phase d'optimisation. Mais avant, voyons comme l'étude de généralisabilité nous permet d'aborder la fiabilité des résultats en fonction de différents projets de mesure.

7.10 Autres projets de mesure

Dans l'exemple qui nous concerne, nous aurions pu chercher à différencier les juges ou les plongeons. Un juge est-il toujours aussi sévère peu importe les sujets ou les plongeons qu'il doit noter ? Les plongeons sont-ils de la même difficulté pour tous les sujets, peu importe le juge qui les note ? Voilà autant de questions légitimes qui font intervenir d'autres plans de mesure.

Supposons que nous souhaitions différencier les juges quant aux points qu'ils accordent. La variance occasionnée par la facette « juges » devient alors une facette de différenciation et les facettes « sujets » et « plongeons » deviennent facettes d'instrumentation. Si la facette « sujets » ou la facette « plongeons » interagit avec la facette juges nous avons autant de sources d'erreur relative. Enfin, si les plongeons à évaluer sont particulièrement difficiles ou les athlètes particulièrement bons, le nombre de points accordés par les juges risque de changer. Ces sources d'erreur absolue s'ajoutent aux sources d'erreur relative.

Le tableau 20 présente le calcul de la généralisabilité pour ce nouveau plan de mesure où il s'agit de trouver la généralisabilité des scores des juges et non celle des sujets. Les résultats de l'analyse de variance demeurent identiques à ceux du tableau

17 parce que les données sont les mêmes. Toutefois, en raison des changements apportés à notre plan de mesure, les composantes de variance diffèrent et sont réparties, comme nous venons de le décrire, entre la variance de différenciation et la variance d'instrumentation (erreur relative ou absolue).

Tableau 20 – Résultats de l'étude G pour le plan de mesure (J/SP)

Source	Variance de différenciation	Source	Variance d'erreur relative	Variance d'erreur absolue
		S		0, 08362
J	1, 72855			
		SJ	0, 02363	0, 02363
		P		0, 09197
		SP		0, 00164
		JP	0, 01388	0, 01388
		SJP	0, 00000	0, 00000
Total (variance)	1, 72855		0, 03752	0, 21474
Écarts types			0, 1937	0, 4634
Coefficient de généralisabilité			0, 979	0, 889

Les résultats du tableau 20 révèlent de très bons coefficients de généralisabilité, que l'on prenne en ligne de compte l'erreur relative (0,979) ou l'erreur absolue (0,889). Dans le premier cas, nous sommes assurés d'un ordre de sévérité très fiable des juges en ce qui concerne le nombre de points accordés. Le juge qui accorde le plus de points le fait de façon constante, peu importe le plongeon ou le plongeur à noter. Enfin, la valeur absolue des points accordés par les juges est également très fiable, peu importe le sujet évalué ou le plongeon. Les juges accordent donc le même nombre de points pour l'ensemble des 12 sujets.

Il y a donc une différence importante entre la généralisabilité des scores des sujets et celle des scores des juges. Ceci n'est pas surprenant considérant que le score de chaque juge est calculé sur 12 sujets, alors que le score de chaque sujet n'est fondé que sur l'appréciation de trois juges seulement.

7.11 Optimisation et étude D

La phase d'optimisation permet d'améliorer la généralisabilité des résultats en apportant des changements au plan d'observation, au plan d'estimation ou au plan de mesure. Nous nous limiterons aux changements qu'il est possible d'apporter au plan d'observation.

Les modèles classiques des scores, dont l'étude de la généralisabilité constitue un prolongement, nous ont appris que la fiabilité des scores s'accroît lorsque l'on augmente le nombre des observations. Ce principe découle des lois de l'estimation statistique : plus notre échantillon est grand, plus l'erreur d'estimation est petite. Il en va de même avec les dispositifs complexes d'observation. Plus une facette comporte de niveaux, plus la variance occasionnée par cette facette dans les résultats sera estimée correctement, car les erreurs aléatoires de mesure dues à l'échantillonnage des niveaux de facette ont tendance à s'annuler lorsque leur nombre devient très grand.

L'examen des résultats de l'étude G nous a conduit aux observations suivantes :

a) pour réduire l'erreur absolue, nous devrions réduire la variance d'erreur occasionnée par les juges ;

b) pour réduire l'erreur relative, nous devrions réduire la variance d'erreur causée par l'interaction « sujets × juges ».

La phase d'optimisation nous permet, à partir de notre connaissance des composantes de variance des différentes facettes, d'estimer l'effet d'un accroissement ou d'une diminution du nombre de niveaux sur la généralisabilité des résultats. Cette procédure est analogue à la formule de Spearman-Brown, quoique beaucoup plus complexe.

Le tableau 21 présente les résultats de la phase d'optimisation de notre simulation. Ce tableau comprend, dans sa partie de gauche, les différentes facettes du plan d'observation, les niveaux traités et les tailles des populations ou univers échantillonnés. Toutes les facettes sont considérées comme ayant été tirées au hasard d'une population de taille infinie. On y retrouve enfin le nombre total des observations (108 = 12 × 3 × 3), les valeurs de généralisabilité absolue et relative. La partie de droite estime les valeurs de ces coefficients pour différents scénarios d'échantillonnage des facettes

Tableau 21 – Étude d'optimisation de l'exemple

Facettes	Niveaux traités	Univers	1	2	3	4	5
S	12	INF	12	12	12	12	12
J	3	INF	3	6	24	12	12
P	3	INF	6	3	3	6	3
Total	108		216	216	288	864	432
ρ_δ	0,898		0,906	0,937	0,970	0,968	0,959
σ_δ			0,323	0,259	0,178	0,183	0,208
ρ_Δ	0,560		0,579	0,691	0,837	0,817	0,782
σ_Δ			0,854	0,670	0,443	0,473	0,053

Le premier scénario consiste à doubler le nombre de niveaux de la facette « plongeons ». En demandant à chaque plongeur de réaliser 6 plongeons plutôt que 3 et en conservant le même nombre de juges, on n'améliore pas significativement la généralisabilité relative, ni la généralisabilité absolue, ainsi que les erreurs relatives et absolues. C'était à prévoir, considérant la faible importance de la facette « plongeons » dans la variance des résultats.

Si nous devions doubler le nombre d'observations, il serait de loin préférable d'engager plus de juges. C'est ce que démontre le scénario 2. Chaque plongeur réaliserait toujours trois plongeons mais verrait sa performance notée par six juges au lieu de trois. Les résultats de l'étude d'optimisation indiquent qu'un accroissement du nombre de juges améliore sensiblement la généralisabilité absolue, de même que la généralisabilité relative, qui était déjà très acceptable. Dans le scénario 2, tout comme dans le scénario 1, le nombre d'observations a été doublé (de 108 à 216). Cette fois-ci l'impact sur la généralisabilité est sensible : la généralisabilité d'erreur absolue passe de 0,560 à 0,691.

Les scénarios 3, 4 et 5 estiment les coefficients de généralisabilité qu'il serait possible d'obtenir en augmentant encore davantage le nombre de juges. Avec douze, les coefficients de généralisabilité relative et absolue laissent entrevoir une fiabilité acceptable des résultats. En éducation et en psychologie, il est parfois coûteux et difficile de compter sur la collaboration d'un aussi grand nombre de personnes compétentes.

Souvent, pour faire face à ce problème, on mettra l'accent sur la formation des juges. En préparant les juges à utiliser de façon rigoureuse des instruments de notation et en établissant des consensus quant à l'interprétation à donner aux différents critères de correction, on diminue grandement l'erreur de mesure et on contribue à améliorer les résultats. Enfin, pour que la tâche soit également répartie, on préférera assigner la moitié des sujets à un groupe de 12 juges et l'autre moitié à un autre groupe de 12 juges, plutôt que de demander à 24 juges d'évaluer tous les sujets. En emboîtant ainsi la correction de certains sujets dans un groupe de juges, on diminue la quantité de travaux à noter pour chaque juge tout en profitant des bénéfices liés à un nombre élevé de juges. Lorsque les deux groupes de juges ne diffèrent pas sensiblement entre eux, ce changement du plan d'observation peut constituer une autre façon d'optimiser la mesure.

8. Conclusion

La fiabilité des résultats est au coeur de nos préoccupations en mesure. Sans fiabilité, les résultats ne peuvent être ni pertinents, ni utiles : la route conduisant à la validité des résultats est coupée. Pourtant, cette qualité essentielle est souvent prise pour acquise ou mal comprise. Justifier l'emploi répété d'un instrument de mesure à partir des seules données sur la cohérence interne des items n'est pas plus approprié que d'utiliser un tournevis pour enfoncer un clou. Il est crucial que l'utilisateur et le constructeur de tests comprennent bien la nature des évidences fournies par les études de fiabilité afin de pouvoir les utiliser au bon moment.

La théorie classique des scores a su s'adapter et évoluer pour répondre à des besoins variés. Lorsque l'échantillon de sujets est modeste, elle demeure la méthode de choix. Grâce aux modèles néoclassiques, il est maintenant possible de calculer des erreurs de mesure différentes pour chaque score. Avec la théorie de la généralisabilité, il est possible d'envisager la fiabilité dans des situations complexes d'observation et pour différents projets de mesure.

Lorsque les échantillons d'items et de sujets sont élevés, les *modèles de réponse aux items* (chapitre 8) peuvent mieux répondre aux besoins des spécialistes : qu'il s'agisse de *calibrer* des banques d'items ou de réaliser des opérations de testing à grande échelle. La multiplication des outils rend encore plus délicat le travail du concepteur et de l'utilisateur de tests. C'est pourquoi il est nécessaire d'approfondir les caractéristiques particulières de chaque modèle d'analyse de la fiabilité. Il n'y a pas de modèles parfaits : il n'y a que des modèles qui rendent compte, plus ou moins bien et plus ou moins utilement, de la nature de nos données.

CHAPITRE 5

LA VALIDITÉ DES RÉSULTATS À UN TEST

1. Le concept de validité

Ces cinquante dernières années, le concept de validité et les méthodes de validation ont profondément évolué. Toutefois, Angoff (1988, p.19) souligne, à juste titre, que si le concept a changé, l'importance que lui accordent les psychométriciens est, elle, restée constante : « *En psychométrie, la validité a toujours été considérée comme le concept le plus fondamental et le plus important* ». Pour les concepteurs comme pour les praticiens, l'essentiel est en effet d'être assuré de mesurer ce qu'ils veulent mesurer, et uniquement cela. La précision de la mesure est certes importante mais elle est inutile si le test n'évalue pas, ou évalue mal, la réalité visée par ses concepteurs. Par conséquent, avant de diffuser un test, les constructeurs ont le devoir de présenter des preuves suffisantes que leur instrument mesure bien ce qu'il prétend mesurer. Comme nous allons le voir en détail dans ce chapitre, ce travail de recueil de preuves est un processus long et complexe, toujours inachevé.

Au début des années 50 (Messick, 1988, pp.18-19), la validité était envisagée de manière relativement morcelée. Ainsi, les *Technical Recommandations* de l'*American Psychological Association* (1954) se limitaient à codifier des types de validité (de contenu, prédictive, concomitante et conceptuelle). La même année, dans la 1ère édition de son ouvrage de référence *Psychological Testing*, Anastasi présentait comme bien distinctes la validité apparente, la validité de contenu, la validité factorielle et la validité empirique. Il faut attendre les années 70 pour qu'un effort important soit réalisé dans le sens d'une intégration des différents types de validité. L'aboutissement de cet effort d'intégration est manifeste dans les *Standards for Educational and Psychological Testing* publiés conjointement par l'*American Psychological Association* et

l'*American Educational Research Association* en 1985. Les *Standards* constituent aujourd'hui une référence incontournable pour les spécialistes de la mesure en psychologie et en éducation. Dans le chapitre qui lui est consacré (pp. 9-18), la validité est présentée comme *« un concept unitaire »* se rapportant non au test lui-même mais aux inférences faites à partir des résultats à celui-ci. Dans cette perspective, il est incorrect de parler de la validité d'un test en général. Seules sont valides les inférences en faveur desquelles suffisamment d'arguments et de données empiriques ont pu être rassemblés. Nous ne pouvons donc pas affirmer, par exemple, qu'un questionnaire évaluant l'anxiété est valide *en général.* Nous pouvons uniquement nous prononcer à propos de la validité de diverses inférences faites à partir des scores à ce questionnaire comme, par exemple, la discrimination de l'anxiété normale et de l'anxiété pathologique, la prédiction de l'intégration dans le milieu professionnel en fonction du degré d'anxiété, l'évaluation de l'efficacité d'un traitement de l'anxiété...

Suivant cette conception de la validité, la validation d'un test est un processus toujours continu d'accumulation de preuves. Les types de validité, définis dans les ouvrages des années 50 et 60, sont aujourd'hui envisagés comme des moyens de validation servant à rassembler des arguments en faveur de telle ou telle inférence. Les *Standards* (1985) distinguent trois grands types de validation (tableau 1) :

(1) La *validation relative au contenu.* Elle consiste à demander à des experts d'évaluer dans quelle mesure les items d'un test sont représentatifs du concept ou du domaine visé. Par exemple, les experts devront apprécier si les items d'un test de définition de mots sont bien des termes appartenant au domaine du français courant. Ou encore, ils devront évaluer si les items d'un questionnaire de dépression représentent bien les différentes facettes du concept de dépression défini par les auteurs de ce questionnaire. Par définition, cette modalité de validation des tests est subjective. Toutefois, si elle respecte une méthodologie rigoureuse, elle permet d'arriver à des conclusions solides qui pourront trouver confirmation dans des recherches empiriques ultérieures.

(2) La *validation en référence à un critère externe.* La procédure de validation repose ici sur l'examen des corrélations entre les scores au test et une autre mesure prise comme critère. Le critère externe peut être de deux types, ce qui donne lieu à deux formes particulières de validation : la *validation concomitante* et la *validation prédictive.* La validation concomitante consiste à évaluer le degré de corrélation entre les scores au test et une mesure prise comme référence. Par exemple, la validité d'un questionnaire conçu pour diagnostiquer la dépression peut être appréciée en comparant les scores à ce test et les évaluations de l'état d'humeur réalisées par des cliniciens expérimentés. La validation prédictive consiste, quant à elle, à évaluer la qualité des prédictions faites sur base des scores au test. Dans ce cas, le critère est la mesure de ce qui a été prédit. Par exemple, la validation d'un test d'admission consistera en la comparaison des scores au test et des résultats obtenus à la fin d'un programme d'études.

(3) La *validation en référence à un concept ou un modèle théorique.* Cette procédure de validation concerne le sens que l'on peut donner aux scores obtenus au test. Tout instrument de mesure repose sur un concept ou un modèle théorique de la réalité que l'on souhaite évaluer. Ce modèle, implicite ou explicite, permet

d'interpréter les données recueillies et de leur donner du sens. Par exemple, dans les théories de la lecture, on distingue aujourd'hui deux procédures intervenant dans la lecture de mots. L'une intervient lorsque le lecteur décode des mots réguliers rencontrés pour la première fois, l'autre fonctionne lorsque le lecteur doit lire des mots irréguliers (p.e. « *femme* »). Si le sujet parvient à lire correctement les mots de la première catégorie mais échoue à lire ceux de la seconde catégorie, ce phénomène pourra être interprété à la lumière du modèle théorique : une des procédures de lecture de mots n'est pas opérationnelle. Pour valider un test de lecture de mots qui s'appuie sur un tel modèle théorique, il est nécessaire de vérifier si les scores au test se conforment aux exigences du modèle. Ainsi, les mots réguliers devront être, en majorité, lus correctement ou incorrectement puisque, selon le modèle, ils font tous appel à une même procédure, laquelle est ou n'est pas fonctionnelle. Si la lecture des mots réguliers est erratique, il faudra alors s'interroger sur la validité des items : pourquoi certains de ceux-ci semblent ne pas mettre en oeuvre les procédures visées ?

À ces trois types de validation, on ajoute habituellement la *validation apparente (face validity).* Elle consiste en une évaluation de surface des items d'un test par des juges. Par exemple, on peut demander aux juges d'évaluer si, à leur avis, les items d'un test semblent bien évaluer les connaissances générales. Les juges ne sont pas nécessairement des experts du domaine et n'ont aucune méthodologie particulière pour effectuer leur travail. Il s'agit, par conséquent, de la procédure de validation la moins scientifique. Elle ne doit pas être confondue avec la validation du contenu qui, elle, fait appel à des juges entraînés appliquant une méthode d'évaluation rigoureusement contrôlée. Le caractère superficiel et peu rigoureux de la validation apparente a entraîné son rejet par de nombreux chercheurs. Toutefois, certains auteurs (Anastasi, 1982, p.136) considèrent qu'elle peut être utile pour mettre au point des instruments destinés à un large public (par exemple, des tests d'admission). Elle permet en effet de créer des tests mieux acceptés par les utilisateurs car leur contenu apparaît plus légitime à ces derniers.

Tableau 1 – Synthèse des différents types de validation d'un test

Type de validation	Caractéristiques
Apparente	Évaluation superficielle de la qualité des items
Contenu	Évaluation formalisée de la qualité des items par des experts
En référence à un critère : concomitante	Évaluation du degré de corrélation des scores à l'item ou à l'échelle avec une mesure prise comme référence
En référence à un critère : prédictive	Évaluation de la prédiction d'une observation future réalisée à partir des scores à l'item ou à l'échelle
Conceptuelle	Évaluation du sens à attribuer aux scores à l'item ou à l'échelle sur base d'un modèle théorique

Ces dernières années, le concept de validité a évolué, à la fois dans le sens d'un élargissement et dans celui d'une plus grande unité. Les travaux de Messick (1988, 1989, 1995) ont joué un rôle important dans cette évolution. Messick souligne que le concept traditionnel de validité est incomplet car il ne prend pas en compte les conséquences sociales de l'usage des tests. Puisque la validité n'est pas une propriété des instruments mais bien des inférences faites à partir des scores, il est nécessaire d'évaluer les conséquences positives et négatives de ces inférences. Dans quelle mesure l'application d'un test se révèle-t-elle utile pour prendre une décision particulière ? N'entraîne-t-elle pas des effets non désirés ? Pour répondre à ces questions, les constructeurs de tests doivent pouvoir anticiper les usages de leurs instruments. Ceci est loin d'être évident a priori. Dès lors, la question de la validité doit être posée pour chaque nouvel usage d'un test.

Par ailleurs, Messick affirme avec force que le concept de validité doit être unifié sous la bannière de la validité conceptuelle. Nous avons vu plus haut que la validité conceptuelle concerne le sens que nous pouvons attribuer aux scores sur base du modèle théorique ayant servi à la conception du test. La question de la signification des scores n'est pas spécifique à une modalité de validation. Elle se pose lorsque nous validons le contenu d'un test et lorsque nous évaluons ses corrélations avec d'autres mesures. Elle se pose également lorsque nous apprécions les conséquences de l'usage des scores. Toutes les informations rassemblées à propos de la signification des scores à un test concourent en fait à répondre à cette question essentielle : *que peut-on dire et faire à partir de ces scores ?*

Dans la suite de ce chapitre, nous allons aborder de manière approfondie la plupart de ces facettes de la validité au travers du découpage classique entre la validité de contenu, la validité en référence à un critère externe et la validité conceptuelle (ou théorique).

2. Validité de contenu

La validation du contenu d'un test consiste à évaluer dans quelle mesure les divers aspects de ce test sont représentatifs du concept visé. Le terme *« aspect »* est utilisé à dessein. Trop souvent, la validation du contenu d'un test est focalisée sur le seul contenu des items. Cet aspect est essentiel mais ne constitue pas la totalité du contenu d'un test. Il est également nécessaire d'évaluer les instructions données aux sujets, les modalités de présentation des stimuli (p.e. présentation papier/crayon, sur écran...), les contraintes de temps, les modalités de réponse (p.e. réponses écrites ouvertes, choix d'images...) et les critères de cotation. Tous ces aspects peuvent concourir à une mesure valide du concept visé. Ils peuvent aussi être la source de biais importants qui détériorent la qualité des mesures réalisées à l'aide du test considéré. Imaginons, par exemple, un test de mémoire, destiné à l'évaluation des troubles de la mémoire chez les personnes âgées, dont les items sont présentés sur écran d'ordinateur en temps limité. La validation de contenu de ce test demandera bien entendu une évaluation du contenu des items. Mais elle demandera aussi une évaluation (1) des consignes données verbalement par le psychologue et par écrit via l'écran, (2) de la présentation des stimuli sur écran, (3) des modalités de réponse à l'aide de la souris et

du clavier, (4) de la limite du temps de réponse, (5) du système de cotation dichotomique « *réussite-échec* ». Tous ces aspects du test concourent-ils bien à une évaluation des troubles de la mémoire ? N'introduisent-ils pas des variables parasites dans la mesure de la mémoire ? Une réponse précise à ces questions est essentielle pour garantir la validité de contenu du test.

Dans la définition proposée ci-dessus, le terme « *concept* » désigne de manière générique ce qui est visé par un test. Un instrument psychométrique ne mesure en effet pas directement une réalité mentale mais, plutôt, une représentation ou un modèle de cette dernière. Par exemple, un test d'intelligence ne permet d'évaluer que le modèle de l'intelligence défendu par son concepteur. D'autres modèles de l'intelligence sont possibles et donneront lieu à des mesures sensiblement différentes. La même remarque peut être faite à propos de tests destinés à l'évaluation de l'estime de soi, de la lecture, de la mémoire... Dans tous les cas, c'est une certaine conception de la réalité qui est évaluée et non une réalité absolue, dégagée de tout point de vue particulier. Les psychométriciens anglo-saxons utilisent le terme « *construct* » pour désigner cette représentation et cette modélisation de la réalité. Aucun terme français ne correspond exactement à « *construct* ». Le terme « *concept* », qui, selon le dictionnaire Robert, désigne une « *représentation mentale générale et abstraite d'un objet* », apparaît comme l'équivalent le plus satisfaisant.

Toute démarche rigoureuse de validation doit débuter par une définition précise du concept visé par le test. La qualité de la validation de contenu dépend étroitement de la précision avec laquelle le concept a été défini et de l'accord des experts à propos de ses facettes. Le terme « *facette* » peut désigner, selon le concept visé, des catégories de comportement (p.e., les divers types de comportements caractéristiques de l'obsession), les composantes d'une compétence cognitive (p.e. les divers traitements intervenant dans le décodage de mots), les capacités intervenant dans une activité professionnelle (p.e. les capacités nécessaires au travail de secrétaire), un ensemble d'objectifs pédagogiques coordonnés (p.e. les objectifs en mathématique de fin de scolarité primaire)... Un concept défini de manière trop floue ne permettra jamais d'obtenir une validité satisfaisante de l'instrument créé pour le mesurer.

Prenons l'exemple de la création d'un questionnaire destiné à diagnostiquer les personnalités schizoïdes. Le DSM-IV (APA, 1994, pp.638-641) présente une définition du concept de « *personnalité schizoïde* » qui est le fruit d'un large consensus entre les cliniciens. A ce titre, cette définition constitue une base solide pour la construction du questionnaire. Selon le DSM-IV, les critères permettant de diagnostiquer une personnalité schizoïde sont :

A. Mode général d'indifférence aux relations sociales et restriction du registre d'expression des émotions en situation interpersonnelle, apparaissant au début de l'âge adulte et présent dans divers contextes, comme en témoignent au moins quatre des manifestations suivantes : (1) ne recherche ni ne prend plaisir aux relations proches, y compris les relations au sein de la famille, (2) choisit presque toujours des activités solitaires, (3) manifeste peu ou pas de désir d'avoir des expériences sexuelles avec une autre personne, (4) prend plaisir à peu ou à aucune activité, (5) n'a pas d'ami ou de confident autres que ses parents au premier degré, (6) apparaît indifférent

aux éloges et aux critiques que lui adressent les autres, (7) manifeste une froideur émotionnelle, du détachement ou une activité limitée.

B. Ne survient pas exclusivement au cours de l'évolution d'une Schizophrénie, d'une Trouble de l'Humeur avec Caractéristiques Psychotiques, d'autres Troubles Psychotiques ou d'un Trouble Envahissant du Développement. N'est pas dû aux effets physiologiques directs de l'état de santé général.

Cette définition nous permet de déterminer les facettes qui devront être prises en compte pour sélectionner les items du questionnaire. Elle nous permet également de préciser les variables qui ne font pas partie du concept. Dans le cas d'un questionnaire clinique, les items sont généralement des affirmations pour lesquelles le sujet doit répondre si elles sont vraies ou fausses pour lui-même (par exemple, *« j'aime la compagnie des autres » vrai - faux*). Des spécialistes du domaine vont générer de tels items susceptibles d'évaluer chacune des facettes du concept. La validation du contenu des items sera ensuite réalisée par un ensemble d'experts qui auront à apparier les items et les facettes (quelle facette est mesurée par un item donné ?). Les experts vérifieront ainsi si les différentes facettes du concept sont bien prises en compte par les items du questionnaire. On demandera également aux experts d'évaluer si des variables parasites n'influencent pas indûment les réponses à certains items (p.e. certains mots de vocabulaire ne risquent-ils pas d'entraîner des erreurs de compréhension des items par des personnes âgées ?). On invitera enfin les experts à évaluer le poids à donner à chacune des facettes au sein du score total au questionnaire. Cette dernière question est importante car la validité d'un score total dépend non seulement de la qualité des scores qui le composent mais aussi de l'importance relative accordée à chacun de ces scores. Quel serait, par exemple, la validité d'un questionnaire de diagnostique de la personnalité schizoïde dont la moitié des items concernerait uniquement le manque d'appétence sexuelle, qui n'est qu'une des facettes du concept de personnalité schizoïde.

Les indices de validité récoltés lors de la validation du contenu sont conditionnels. Ils dépendent en effet de la définition du concept visé. Cette définition peut être remise en cause. Par exemple, la définition du concept de personnalité schizoïde peut changer en fonction de l'évolution des connaissances en psychologie clinique. Par conséquent, certaines facettes mesurées par le questionnaire peuvent, à un moment donné, devenir inadéquates. Les indices de validité sont également relatifs à la fonction assignée au test. Par exemple, les experts peuvent considérer comme valide le contenu d'un test de mathématique destiné à servir d'examen d'admission. Par contre, ce même contenu peut être considéré comme peu valide si le test doit servir à diagnostiquer des difficultés d'apprentissage en mathématique. Enfin, les indices de validité dépendent de la population visée par le test. Les experts ne jugeront pas la validité de contenu d'un test de lecture de la même manière si celui-ci est destiné à des élèves belges ou à des élèves québécois. En effet, d'un pays à l'autre, le curriculum d'étude et la familiarité avec le vocabulaire peuvent différer sensiblement. Un élément de validité pour les uns peut être une source de biais pour les autres. Le caractère conditionnel de la validité de contenu implique que celle-ci ne peut être tenue pour acquise une fois pour toutes. La validation du contenu d'un test reste toujours relative au temps et au lieu où elle a été réalisée. Elle doit, par conséquent, être réévaluée périodiquement.

Haynes, Richard et Kubany (1995, pp 244-247) proposent une synthèse très utile des règles de base qui devraient être suivies lors de la validation du contenu d'un test. Le tableau 2 présente les sept règles essentielles que devrait respecter tout constructeur de test soucieux de produire un instrument valide.

Tableau 2 – Principes de base pour la validation du contenu d'un test (d'après Haynes et al., 1995)

1.	**Définir avec soin le domaine et les facettes du concept et valider cette définition.**
2.	**Utiliser un échantillon d'experts et de membres de la population de référence pour créer les items et les autres aspects du test.**
3.	**Soumettre tous les aspects du test à une validation de contenu.**
4.	**Utiliser plusieurs experts pour valider le contenu d'un test et quantifier leurs jugements à l'aide d'échelles formalisées.**
5.	**Examiner la représentation proportionnelle des items relativement aux différentes facettes du concept.**
6.	**Présenter les résultats de la validation de contenu lors de la publication de tout nouvel l'instrument.**
7.	**Prendre en compte toutes les analyses psychométriques ultérieures pour affiner la validation du contenu du test.**

Le jugement des experts joue un rôle crucial dans la procédure de validation du contenu d'un test. Les principes de validation 4 et 5 (tableau 2) impliquent que ces jugements soient quantifiés. Dans la suite de cette section, nous allons présenter plusieurs indicateurs quantitatifs de validité couramment utilisés lors de la mise au point de tests.

Dans le domaine de l'éducation, Crocker et Algina (1985) énumèrent cinq indicateurs utilisés pour évaluer dans quelle mesure un ensemble d'items sont représentatifs des objectifs pédagogiques visés par le test :

(1) le pourcentage d'items appariés aux objectifs ;

(2) le pourcentage d'items appariés aux objectifs jugés très importants ;

(3) la corrélation entre le poids des objectifs et le nombre d'items les mesurant (Klein et Kosecoff 1975) ;

(4) l'indice de congruence item-objectif (Hambleton, 1980) ;

(5) le pourcentage des objectifs non mesurés par les items.

Ces cinq catégories d'indices ne fournissent pas une information équivalente sur la congruence item-objectif. Les deux premiers, en particulier, nécessitent un échantillonnage important d'items. Enfin, le troisième ne fournit pas de résultats intéressants si tous les objectifs sont d'égale ou à peu près d'égale importance. Une faible variation des valeurs de pondération de chaque objectif entraînera une diminution de la valeur maximale de la corrélation.

Crocker et Algina (1985) ont proposé une version simplifiée de l'indice de Hambleton. L'indice de congruence de l'item i à l'objectif k est calculé par la formule suivante :

$$I_{ik} = \frac{N}{2N-2}(\bar{X}_k - \bar{X}) \quad (5.1)$$

N = le nombre d'objectifs,

$\bar{X}$ = la moyenne des évaluations de l'item i pour tous les objectifs,

$\bar{X}_k$ = la moyenne des évaluations de l'item i pour l'objectif k.

L'indice I varie de -1 à +1 ; la valeur de 1 n'étant possible que lorsque tous les juges ont apparié chaque item à un seul et même objectif. Le tableau 3 présente un exemple de calcul de l'indice de congruence de Hambleton tel que simplifié par Crocker et Algina (1985). Les calculs sont effectués pour les résultats de trois juges évaluant sept items par rapport à trois objectifs. Les juges (J1 à J3) ont eu à se prononcer sur la congruence entre chaque item (1 à 7) et chaque objectif (objectif 1 à objectif 3) : +1 indique que l'item mesure l'objectif, -1 qu'il ne le mesure pas et 0 indique que le juge est incertain.

Deux valeurs essentielles sont calculées (en italique dans le tableau) pour chaque item : la moyenne par objectif des évaluations des juges pour chaque item ($\bar{X}_1$ à $\bar{X}_3$), ainsi que la moyenne, pour l'ensemble des objectifs, des évaluations des juges pour chaque item ($\bar{X}$) . La deuxième partie du tableau fournit les valeurs de l'indice de congruence calculé par la formule 5.1 pour chaque paire item-objectif. On peut remarquer que, selon les trois juges interrogés, l'objectif 1 est mesuré principalement (valeurs de I en gras) par les items 2, 6 et 7, l'objectif 2, par les items 1 et 4 et l'objectif 3 par les items 3 et 5. L'item 7 est le seul à démontrer une congruence parfaite. En effet, tous les juges se sont accordés pour affirmer qu'il mesurait l'objectif 1 et qu'il ne mesurait ni l'objectif 2, ni l'objectif 3. C'est pourquoi il se mérite la valeur maximale de 1.

Soulignons qu'il est important d'utiliser plus d'un seul indice de congruence. Il est en effet plus facile d'obtenir un indice I élevé lorsque le calcul de l'indice ne porte que sur un nombre réduit des objectifs à couvrir. Le pourcentage des objectifs non mesurés par les items devrait, par conséquent, toujours accompagner l'indice I pour mieux saisir la portée de ce dernier.

Jusqu'à quel point peut-on compter sur le jugement des experts pour évaluer la validité de contenu d'un test ? A cet égard, l'indice de Hambleton (1980) ne fait que calculer la congruence entre item et objectif sans tenir compte du fait qu'un ou plusieurs juges peuvent ne pas concorder dans leur appréciation avec les autres juges. Le degré de consensus (ou de fiabilité) entre les juges peut être évalué par trois indicateurs :

1. *La variance des jugements* : il n'y a que peu de dispersion parmi les évaluations effectuées pour un même item. Les juges ont tendance à attribuer la même cote à un même item.

2. *La concordance des jugements* : les juges ont tendance à ordonner de la même manière les items selon leur degré de congruence avec la facette à mesurer. L'item le plus congruent pour un juge est également le plus congruent pour les autres juges.
3. *La cohérence interne des jugements* : les juges sont consistants dans leur manière d'évaluer les items par rapport aux autres juges. Un juge sévère demeure sévère pour tous les items, et non pas seulement pour quelques-uns d'entre eux.

Tableau 3 – Illustration du calcul de l'indice de congruence items/objectifs

	Objectif 1				**Objectif 2**				**Objectif 3**				
Items	J1	J2	J3	$\bar{X}_1$	J1	J2	J3	$\bar{X}_2$	J1	J2	J3	$\bar{X}_3$	$\bar{X}$
1	-1	0	0	*-0,33*	-1	1	1	*1*	-1	0	-1	*-0,67*	*0*
2	1	1	1	*1*	1	-1	-1	*-1*	1	1	0	*0,33*	*0,11*
3	0	1	1	*0,67*	0	0	-1	*-0,67*	0	1	1	*1*	*0,33*
4	-1	-1	-1	*-1*	-1	1	1	*0,67*	-1	-1	1	*-0,33*	*-0,22*
5	0	-1	-1	*-0,67*	0	0	-1	*-0,67*	0	0	1	*0,67*	*-0,22*
6	1	1	0	*0,67*	1	-1	0	*-0,33*	1	-1	-1	*-0,67*	*-0,11*
7	1	1	1	*1*	1	-1	-1	*-1*	1	-1	-1	*-1*	*-0,33*

	Indices I pour les trois objectifs		
Items	I_1	I_2	I_3
1	-0,25	**0,75**	0,50
2	**0,67**	-0,83	0,17
3	0,26	-0,75	0,50
4	-0,59	**0,67**	**-0,08**
5	-0,34	-0,34	0,67
6	**0,59**	-0,17	-0,42
7	**1,00**	0,50	-0,50

Supposons que nous ayons demandé à un groupe de juges d'apprécier, sur une échelle de 1 à 5, dans quelle mesure une série de questions évalue bien une des facettes

d'une personnalité donnée. Plus la moyenne des évaluations de chaque question est élevée, plus cette question est jugée pertinente par les juges. La figure 1 illustre comment représenter graphiquement l'indicateur de variance dans cette situation. Chaque point de ce diagramme de dispersion est déterminé par les coordonnées suivantes :

1. en abscisse, la valeur moyenne des évaluations des juges concernant la pertinence d'une question ;
2. en ordonnée, l'écart type de la distribution des évaluations concernant cette même question.

Il ne suffit pas qu'un item reçoive une évaluation moyenne élevée pour juger de sa pertinence. Cette évaluation doit aussi être sensiblement la même pour un grand nombre de juges. Par exemple, si deux items reçoivent une cote moyenne de 4, celui dont la variance des jugements est égale à 0,6 possède une meilleure validité de contenu que celui dont la variance est égale à 1,4. Dans le diagramme de dispersion de la figure 1, les items possédant la meilleure validité de contenu se situent, par conséquent, dans le quatrième quadrant.

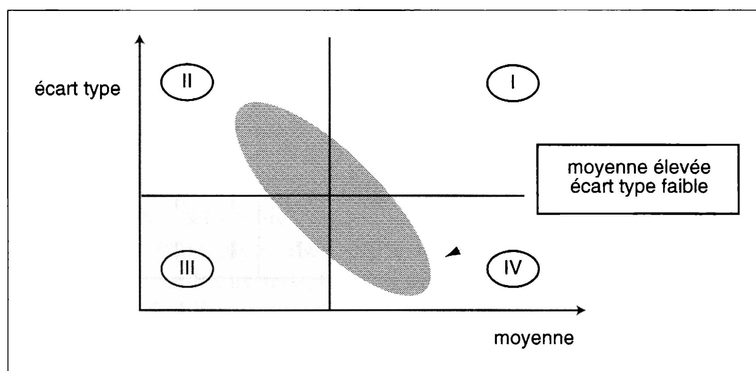


Figure 1 – Relation entre moyenne et écart type des jugements concernant des jugements concernant un ensemble d'items

L'évaluation de la concordance des jugements est un moyen de vérifier la fiabilité des jugements des experts appelés à se prononcer sur la validité de contenu des items. Dans ce cas, l'attention se porte sur le classement des items réalisé par le juges. Seule la place de l'item dans le classement des juges est ici prise en compte. Le score attribué à l'item n'est pas considéré. Supposons, par exemple, que l'item 6 soit jugé par tous les experts comme celui qui mesure le mieux un trait de personnalité. Cette concordance des jugements signifie que cet item a reçu la cote la plus élevée donnée par chaque expert. Sur une échelle de 1 à 5, cette valeur peut être 3 pour un expert, 4 pour un autre et 5 pour un troisième. Malgré cette différence de scores, l'item 6 est évalué de manière concordante par les trois juges puisque ceux-ci lui accordent tous leur score le plus élevé.

Le *coefficient W de Kendall* (1948) permet de mesurer le degré de concordance entre plusieurs juges. Cet indice complète bien l'indice de variance présenté ci-dessus car une part de la dispersion des résultats entre les juges peut provenir de la manière

dont ils utilisent l'échelle d'évaluation. Certains juges ont tendance à polariser leurs opinions et à n'employer que les valeurs extrêmes (p.e. 1 ou 5). D'autres, au contraire, situent leurs appréciations près du centre et évitent les valeurs extrêmes (p.e. 2, 3 ou 4). Ces différentes pratiques influencent la dispersion des appréciations des juges et donc l'indice basé sur la variance. Par contre, elles n'ont pas d'impact sur le classement des items. Plusieurs juges peuvent avoir le même classement des items alors que la variance de leurs évaluations est différente.

Le calcul de la valeur du W de Kendall se fait en trois étapes :

1. transformer les scores observés en rangs pour chaque juge (tableau 4) ;
2. calculer la valeur de s (tableau 5) ;
3. calculer la valeur de W et son degré de signification (équation 5.3).

La première étape est la plus simple. Il s'agit d'ordonner les N items pour chacun des k juges. Si les juges concordent entre eux, comme c'est le cas de l'exemple du tableau 4, l'ordre de leur appréciation devrait être le même pour tous et se traduire par des rangs semblables.

Tableau 4 – W de Kendall : transformation en rangs

	Score donné à l'item			Rang de l'item		
	item 1	item 2	item 3	item 1	item 2	item 3
juge 1	3	5	2	2	3	1
juge 2	2	3	1	2	3	1
juge 3	1	2	0	2	3	1

Une fois la transformation effectuée, il faut ensuite calculer la valeur de s. Cette valeur est égale à la somme des écarts entre la somme des rangs attribués à chaque item et la moyenne de la somme des rangs pour tous les items, le tout élevé au carré (formule 5.2). Plus la somme des écarts est grande, plus les juges concordent. En effet, s'ils ne concordaient pas, la somme des rangs pour chaque item serait approximativement la même et ne différerait pas de la moyenne.

$$s = \sum\left(R_j - \sum\frac{R_j}{N}\right)^2 \qquad (5.2)$$

R_i = somme des rangs accordés à l'item j

N = nombre d'items

Le tableau 5 présente un exemple de calcul de la valeur de s pour les données du tableau 4. Cette valeur est ensuite utilisée dans l'équation de calcul du W de Kendall pour trouver la valeur du coefficient. En voici la formule :

$$W = \frac{s}{\frac{1}{12}k^2(N^3 - N)} \qquad (5.3)$$

Tableau 5 – W de Kendall : calcul de s

	item 1	item 2	item 3	
juge 1	2	3	1	
juge 2	2	3	1	
juge 3	2	3	1	
R_j	6	9	3	
$\sum \frac{R_j}{N}$	6	6	6	$s = \sum (6-6)^2 + (9-6)^2 + (3-6)^2 = 18$

En fait, il s'agit de diviser *s* par la valeur maximum que *s* peut prendre avec *k* juges et *N* items (expression au dénominateur). Dans notre exemple, la valeur de *s* est égale à la valeur maximum et le coefficient de Kendall est dès lors égal à 1. Cette valeur correspond à une concordance parfaite des classements effectués par les différents juges. Voici le détail des calculs de la valeur de *W* pour notre exemple :

$$W = \frac{18}{\frac{1}{12}(3^2(3^3 - 3))} = \frac{18}{18} = 1$$

On peut enfin vérifier si la valeur de *W* est significativement différente de 0. Lorsque $N>7$, il est possible de transformer la valeur de *W* en valeur de χ^2 se distribuant avec $N-1$ degrés de liberté. La transformation s'effectue à partir de l'équation suivante :

$$\chi^2 = \kappa(N-1)W \qquad (5.4)$$

Une mesure alternative de l'accord entre les juges est donnée par *le coefficient* κ *(kappa) de Cohen.* Ce coefficient postule que les données soient nominales. Ce coefficient est, par conséquent, indiqué lorsque la tâche demandée aux juges est un classement des items dans des catégories. Par exemple, les juges peuvent être invités à mettre en correspondance des affirmations (p.e. *« j'aime les fleurs » ; « j'apprécie le travail en groupe »...*) et différentes facettes de la personnalité qu'elles sont sensées représenter (p.e. *« introversion » ; « extraversion »...).* Dans ce cas, les facettes de la personnalité sont prises comme des catégories au sein desquelles les items doivent être rangés.

Le coefficient κ prend en compte le nombre de fois où les juges sont d'accord mais prend également en compte le nombre d'accords simplement dus au hasard. Par conséquent, ce coefficient est plus exigent que la plupart des autres indices de concordance et sera habituellement plus faible que ceux-ci. Le coefficient κ est le rapport entre la proportion de fois où les juges sont d'accord (corrigée pour accords dus à la

chance) et la proportion maximum de fois où ceux-ci pourraient être d'accord (également corrigée pour accords dus à la chance) :

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \tag{5.5}$$

$P(A)$ = proportion de fois où les juges sont d'accord,

$P(E)$ = proportion de fois où l'on s'attend à ce que les juges soient d'accord uniquement par chance.

La valeur de κ est égale à 1 s'il y a un accord parfait entre les juges. Si, par contre, les accords ne dépassent pas ceux qui étaient attendus du fait de la chance, la valeur de κ est égale à 0.

Nous allons illustrer le calcul de κ avec un exemple simple où quatre juges ont chacun à classer quatre items en trois catégories. Partant des données présentées dans le tableau 6, nous pouvons calculer $P(E)$ à l'aide de la formule suivante :

$$P(E) = \sum\left(\frac{C_j}{Nk}\right)^2 \tag{5.6}$$

C_j = somme des fréquences de la catégorie j

N = nombre d'items

k = nombre de juges

Dans notre exemple, $P(E) = \left(\frac{5}{16}\right)^2 + \left(\frac{5}{16}\right)^2 + \left(\frac{6}{16}\right)^2 = 0,336$

Tableau 6 – Exemple de calcul de coefficient κ de Cohen

	catégorie 1	catégorie 2	catégorie 3	s_i
item 1	4	0	0	12/12 = 1
item 2	1	3	0	6/12 = 0,5
item 3	0	2	2	4/12 = 0,333
item 4	0	0	4	12/12 = 1
C_j	5	5	6	

Avant de calculer $P(A)$, il est nécessaire de calculer s_i pour chacun des items au moyen de la formule suivante :

$$s_i = \frac{1}{k(k-1)}\sum n_{ij}(n_{ij} - 1) \tag{5.7}$$

n_{ij} = fréquence de l'item i dans la catégorie j

k = nombre de juges

Les valeurs de sont présentées dans la dernière colonne du tableau 6. Une fois ces valeurs calculées, on peut déterminer la valeur de *P(A)* au moyen de la formule suivante :

$$P(A) = \frac{1}{N}\sum s_i \tag{5.8}$$

Dans l'exemple, $P(A) = \frac{1}{4}(1 + 0,5 + 0,333 + 1) = 0,708$

Nous pouvons alors calculer la valeur de $\kappa = \frac{0,708 - 0,336}{1 - 0,336} = 0,560$. Cette valeur nous renseigne sur l'existence d'un accord modéré entre les quatre juges à propos du classement des quatre items.

La cohérence interne est le dernier des trois indices utiles pour apprécier de manière quantitative les jugements des experts à propos de la validité de contenu. La signification et les méthodes de calcul de la cohérence interne ont été abordés dans le chapitre 4. Il se s'agit ni d'une valeur de concordance ni du degré d'accord entre les juges. Elle nous permet plutôt de déterminer si les juges sont constants dans leurs jugements. Ainsi, un juge sévère dans son appréciation d'un item devrait l'être pour tous les autres items, et réciproquement.

L'évaluation de la dispersion, de la concordance inter-juges et de la cohérence interne des appréciations nous fournit des indices différents, mais complémentaires, du degré de confiance que l'on peut avoir dans l'évaluation de la validité de contenu d'un test par un groupe de juges. Les items les plus valides seront ceux pour lesquels les juges auront manifesté le moins de dispersion dans leur appréciation, la plus grande concordance dans leur classement respectif et la meilleure constance d'un item à l'autre. Ces outils statistiques nous aident à mieux comprendre la différence qui existe dans la pratique entre validité apparente et validité de contenu.

3. Validité en référence à un critère externe

3.1 PRINCIPES GÉNÉRAUX

Si un test mesure une caractéristique particulière, il devrait être bien corrélé avec tout critère mesurant la même caractéristique ou une caractéristique voisine. Le test et le critère devraient donc partager une part importante de variance commune. Pour démontrer la validité des résultats d'un test, le constructeur peut faire appel à deux types de critère :

1. Le critère le plus facile à trouver est sans doute une autre mesure dont la validité est reconnue et qui a déjà fait ses preuves. Une forte corrélation entre le test et le critère externe permet de penser que nous avons affaire à deux mesures de la même caractéristique ou du même trait. C'est la *validité concomitante.*
2. Le critère peut aussi être un indicateur d'une performance que l'on cherche à prédire. C'est la *validité prédictive.*

Parfois, le critère est relativement simple à mesurer. Par exemple, la grandeur d'un enfant à quatre ans peut être un bon prédicteur de la grandeur à l'âge adulte. Pour

le démontrer, il suffit de mesurer la grandeur de plusieurs sujets choisis au hasard à l'âge de quatre ans (prédicteur) et de prendre de nouveau la même mesure à l'âge adulte (critère).

Parfois, le critère paraît simple à mesurer, mais les apparences peuvent être trompeuses. C'est le cas par exemple du décrochage scolaire. Un test de dépistage du décrochage scolaire posséderait une bonne validité prédictive s'il y avait une forte corrélation entre le résultat au test et le décrochage futur de l'élève. Il faudrait, bien entendu, définir opérationnellement ce que l'on entend par « *décrochage scolaire* ». Une telle définition opérationnelle du décrochage scolaire pourrait être :

> Abandon volontaire et prolongé des études, pour une période consécutive d'au moins deux ans, qui n'est ni la conséquence d'une maladie, ni la conséquence d'une sanction de l'institution scolaire.

Selon la définition précédente, une maternité adolescente ferait-elle partie des conditions acceptables de décrochage ? Ce n'est pas une maladie. Ce genre d'abandon peut-il être considéré comme un abandon volontaire ? Pourtant, chez les sujets féminins, c'est un facteur important de décrochage. Si rien n'est fait pour tenir compte de ce facteur dans l'instrument de mesure, le risque est grand que la validité prédictive du test soit meilleure pour les garçons que pour les filles.

Enfin, le critère peut être fort complexe et faire intervenir plusieurs habiletés ou attitudes différentes. Prenons, par exemple, le critère *leadership*. Si nous voulons construire un test qui prédira les capacités de leadership d'un individu, il faut pouvoir mesurer tous les aspects de cette caractéristique. Le critère pourrait être constitué de l'un ou de plusieurs des indicateurs suivants :

- le rapport subjectif des gens qui travaillent sous la direction de l'individu ;
- le rapport subjectif des supérieurs hiérarchiques ;
- l'observation discrète de comportements de leader dans l'exécution d'une tâche.

Le choix et la mesure d'un bon critère peut être une tâche tout aussi problématique que la construction de l'instrument de mesure lui-même. C'est pourquoi elle requière un soin tout particulier. Une étude de validité qui chercherait à établir une corrélation entre les résultats à un test et un critère mal défini au départ pourrait fort bien constituer une perte de temps ou encore conduire au rejet d'un bon instrument de mesure faute de critère adéquat. La définition opérationnelle du critère est l'une des plus importantes considérations pratiques dans l'estimation de la validité liée à un critère externe. D'autres facteurs sont également susceptibles d'influencer l'estimation de cette validité :

- la grandeur de l'échantillon de sujets ayant participé à l'étude de validation ;
- les limites imposées à la dispersion des résultats ;
- la précision (fiabilité) du prédicteur et du critère.

La validité étant souvent calculée au moyen d'une corrélation de Pearson, elle partage également avec celle-ci les mêmes limites et postulats (homocédasticité, normalité des distributions du prédicteur et du critère ; voir chapitre 2). Les deux premiers facteurs de la liste précédente sont des conséquences de ce lien direct.

3.2 Effet de la grandeur de l'échantillon

La validité des résultats d'un instrument de mesure est une estimation plus ou moins entachée d'erreur. En effet, il n'y a aucune certitude quant à la probabilité de retrouver la même valeur de validité avec un échantillon semblable tiré de la même population d'intérêt. La probabilité d'obtenir une valeur stable de validité s'accroît, cependant, lorsque celle-ci est calculée à partir d'un nombre suffisamment grands de résultats. Schmidt, Hunter et Urry (1976) ont démontré qu'avec des échantillons de 200 sujets et plus, la valeur calculée de la validité était celle de la population dans 90% des cas. Cette probabilité diminue à 25% et 35% lorsque l'échantillon n'est que de 30 ou 50 sujets respectivement. Sauf s'il existe une très forte relation entre le prédicteur et le critère, il est par conséquent préférable d'effectuer une étude de validité avec un grand nombre de sujets.

Lorsqu'il est difficile d'effectuer une étude de validité avec de grands échantillons, il faut alors réaliser plusieurs études de validité afin de voir si la corrélation entre le prédicteur et le critère se généralise à un ensemble de situations semblables. Cette trans-validation (« *cross-validation* ») permet d'estimer l'impact des fluctuations d'échantillonnage sur la stabilité de l'estimation de la validité. Cette procédure consiste à calculer la meilleure équation de régression (voir chapitre 2) sur un échantillon et à voir comment elle permet de prédire les résultats d'un autre échantillon tiré de la même population. Deux échantillons ne sont pas toujours nécessaires. Lorsque le nombre de répondants est assez grand, on peut simplement répartir au hasard les sujets de l'échantillon total en deux groupes et calculer une régression linéaire sur chaque moitié.

3.3 Effet de la réduction de l'étendue

Puisque l'estimation de la validité repose très souvent sur le calcul de corrélations, la réduction de l'étendue a les mêmes effets que lors de l'estimation de la corrélation (voir chapitre 2). Celle-ci peut survenir dans trois cas particuliers d'études de validité :

1. *Le test est utilisé pour des fins de sélection*, comme, par exemple, lors d'une demande d'emploi. Si, après avoir sélectionné un groupe d'individus sur la base de leur performance au test, on cherche par la suite à démontrer la validité de l'instrument au moyen d'une corrélation entre les résultats au test et la performance professionnelle, il faut tenir compte que la corrélation ainsi calculée ne comprend plus les valeurs les plus faibles au prédicteur. Cette situation risque d'entraîner la sous-estimation de la véritable validité du test puisqu'elle ne porte que sur une partie de l'échantillon de départ : les candidats qui ont été acceptés.
2. *Le test prédicteur est corrélé avec une variable intervenant dans la sélection des sujets.* C'est le cas lorsque nous cherchons à vérifier la validité prédictive d'un test d'aptitude aux études universitaires. Il est fort peu probable que tous les individus ayant répondu à un tel questionnaire terminent des études universitaires. En effet, les universités possèdent leurs propres politiques d'admission souvent fondées sur les résultats académiques antérieurs du candidat. Les résultats académiques peuvent être en forte corrélation avec le test prédicteur,

puisqu'après tout, ils cherchent à prédire la même chose. D'autres facteurs feront, par exemple, que de bons étudiants ne termineront pas leurs études : difficultés financières, changement d'orientation, etc... Tous ces facteurs font que l'échantillon sur lequel sera calculée la corrélation entre test prédicteur et critère de réussite universitaire (p.e. la moyenne cumulative), n'est pas le même que celui qui s'est présenté au test.

3. *Le test prédicteur peut être trop facile ou trop difficile.* Un test trop facile ne permet pas de différencier suffisamment les élèves forts au test prédicteur et réduit par conséquent la variance des résultats. La même observation vaut également pour un test trop difficile. Les conséquences sont alors les mêmes que celles qui découlent de la réduction de l'étendue.

3.4 EFFET DE LA FIABILITÉ DU PRÉDICTEUR ET DU CRITÈRE

Lorsque nous calculons la validité des résultats à un test, nous réalisons nos calculs sur les valeurs observées du prédicteur et du critère. Ces valeurs sont imprécises, à moins qu'elles n'aient une fiabilité parfaite. Considérant qu'une partie des valeurs observées est constituée d'erreurs aléatoires, il est normal que nous tenions compte de cette erreur dans le calcul de la validité.

Si l'on souhaite estimer la validité, non pas à partir des scores observés, mais à partir des scores vrais, il est nécessaire d'effectuer la *correction dite d'atténuation*, formulée dans l'équation suivante :

$$r_{v_X v_Y} = \frac{r_{XY}}{\sqrt{r_{XX'}}\sqrt{r_{YY'}}} \qquad (5.9)$$

Dans cette équation, le numérateur représente la corrélation entre les scores observés et le dénominateur représente le produit des racines carrées de la fiabilité du prédicteur et du critère. Le résultat de la division nous donne la corrélation corrigée pour atténuation. La correction pour atténuation nous permet d'estimer le *potentiel* de validité d'un test. En effet, si la corrélation corrigée pour atténuation est faible, il y a peu d'espoir d'améliorer la validité du test. C'est sans doute qu'il n'y a pas d'association très forte entre le prédicteur et le critère, dans les conditions où a été réalisée l'étude de validité. Par contre, si la validité, une fois corrigée pour atténuation est beaucoup plus élevée, ceci peut vouloir dire que nous pourrions accroître sensiblement celle-ci en améliorant la fiabilité, notamment en augmentant le nombre d'items.

Supposons que nous ayons obtenu une corrélation de 0,45 entre un prédicteur et un critère. La fiabilité du test prédicteur est de 0,55 et la fiabilité du critère est de 0,70. La corrélation corrigée pour atténuation sera la suivante :

$$r_{v_X v_Y} = \frac{0{,}45}{\sqrt{0{,}55}\sqrt{0{,}70}} = 0{,}73$$

La valeur de 0,73 est la valeur maximale du coefficient de corrélation que nous pourrions obtenir entre scores observés au prédicteur et au critère en postulant qu'il n'y a aucune erreur de mesure. Il s'agirait là d'une très bonne valeur de validité, mais est-il possible de l'atteindre en pratique ? Nous avons souvent peu de prise sur le cri-

tère. Il peut, par conséquent, être difficile d'accroître la précision de cette mesure. Par contre, nous pouvons accroître la fiabilité de notre test prédicteur en augmentant le nombre d'items. Quel effet sur la validité du test aurait l'augmentation du nombre d'items ?

Supposons que nous allions jusqu'à doubler le nombre d'items du test prédicteur. La formule de Spearman-Brown (4.34) nous permet d'espérer la fidélité suivante :

$$\hat{r}_{XX'} = \frac{2r_{XX'}}{1 + r_{XX'}} = \frac{2\,(0{,}55)}{1 + 0{,}55} = 0{,}71$$

Si la fidélité du test devait passer de 0,55 à 0,71, on peut s'attendre à un accroissement significatif de la corrélation entre scores observés. En utilisant cette nouvelle valeur, nous pouvons résoudre l'équation (5.5) pour trouver :

$$\frac{r_{XY}}{\sqrt{0,71}\sqrt{0,70}} = 0,73$$

$$r_{XY} = 0{,}73\,\sqrt{0{,}71}\,\sqrt{0{,}70} = 0{,}51$$

En doublant le nombre d'items et en postulant que la corrélation entre les scores vrais du test prédicteur et du critère demeure la même (0,73), on peut s'attendre à ce que la validité du test passe de 0,45 à 0,51. Cette validité est acceptable, selon les situations, mais elle démontre aussi combien il y a loin de la coupe aux lèvres, c'est-à-dire entre la validité potentielle entre scores vrais et la validité qu'il est possible d'obtenir dans la réalité en améliorant la fiabilité du test prédicteur.

4. Validité conceptuelle (ou théorique)

4.1 Principes généraux

La validité conceptuelle repose sur une collection d'évidences concernant la nature d'un instrument de mesure. À cet égard, c'est probablement le type de validité qui exige la plus grande quantité d'efforts au niveau de la validation des résultats d'un instrument de mesure.

Lorsque l'on traite de validité conceptuelle, l'on ne peut s'empêcher de faire allusion au fait que, dans l'évolution de toute science, la compréhension d'un phénomène va de paire avec notre capacité de le mesurer adéquatement. Qu'il s'agisse de variables composites comme l'intelligence, la motivation scolaire ou les styles cognitifs, notre capacité à tester ces variables et à les étudier dépend de notre habileté à les mesurer. Sans instrument valide de mesure de ces concepts, il est difficile d'entrevoir comment la connaissance et la compréhension de leur rôle dans les phénomènes étudiés peuvent progresser. En retour, sans étude de ces phénomènes et sans une compréhension suffisante, il est difficile de développer des instruments de mesure adéquats.

La validité conceptuelle est donc au coeur du problème de l'opérationnalisation des variables. Pour réaliser une étude de validité conceptuelle, il faut recueillir une grande quantité d'informations. Celles-ci devront découler des prédictions, hypothèses

que l'on peut tirer de la théorie. Lorsque ces hypothèses ne se vérifient pas, deux explications sont possibles :

1. l'instrument de mesure est une bonne opérationnalisation de la théorie, mais l'hypothèse déduite ne se vérifie pas. Il faut alors changer la théorie.
2. la théorie est essentiellement valide, mais l'instrument en est une mauvaise opérationnalisation. Il faut alors revoir l'instrument ou la procédure de validation employée. Ou l'instrument utilisé n'est pas adéquat, ou l'information que l'on recherche n'est pas pertinente à ce que nous voulons prouver.

Prenons le cas d'un instrument de mesure de l'intelligence. L'étude de validité conceptuelle ne sera pas la même pour un test de quotient intellectuel (QI) que pour un test d'intelligence opératoire inspiré de la théorie génétique de Jean Piaget. Les deux théories reposant sur des postulats différents concernant l'intelligence, la construction d'instruments de mesure se fera différemment.

Dans le cas des tests de QI, on s'attend à ce que ce genre de test différencie entre plusieurs types d'intelligence, notamment entre intelligence verbale et non verbale. Les tests appartenant à la catégorie « intelligence verbale » devraient être corrélés plus fortement entre eux qu'avec tout autre test mesurant la catégorie « intelligence non verbale ». Un test d'intelligence verbale qui aurait une corrélation plus forte avec un test d'intelligence non verbale soulèverait des questions, soit quant à la nature de la variable qu'il mesure, soit quant à la façon de la mesurer.

D'autres informations entrent en considération dans le calcul de la validité conceptuelle. Si les résultats des études neurologiques portant sur la spécialisation hémisphérique tendent à indiquer que les hommes réussissent mieux dans le domaine des habiletés spatiales et que les femmes réussissent mieux dans le domaine des habiletés verbales, les résultats aux tests d'intelligence devraient normalement favoriser les femmes dans la catégorie *« intelligence verbale »* et favoriser les hommes dans la catégorie *« intelligence non verbale »*. De tels résultats seraient une confirmation des études neurologiques en plus de constituer une information supplémentaire à l'appui de la validité conceptuelle du test. Dans ce cas-ci, la validité conceptuelle de l'instrument de mesure est démontrée par une différence significative entre deux groupes.

Dans le cas des tests piagétiens, on s'attend à ce que l'ordre de réussite des items soit conforme à la progression développementale des habiletés. Par exemple, des items nécessitant les opérations concrètes devraient être réussis avant les items portant sur des opérations formelles. Il y aurait un problème de validité conceptuelle si des items d'habiletés formelles étaient réussis, alors que des items d'habiletés concrètes du même domaine étaient échoués. On s'attend donc à ce que les items d'un test opératoire constituent une *échelle hiérarchique*. La présence d'une hiérarchie dans l'ordre de réussite des items appuierait la validité conceptuelle du test. C'est pourquoi plusieurs constructeurs de tests opératoires utilisent le calcul d'un *coefficient de reproductibilité* comme indice de la validité conceptuelle de leur test. Un tel indicateur ne serait pas approprié dans une étude de validité de tests de QI parce qu'il n'y a pas dans les théories factorielles de l'intelligence de telles hypothèses sur l'ordre de réussite des items.

L'exemple du tableau 7 illustre comment les résultats à un test peuvent former une hiérarchie. Le tableau de données met en ordre les sujets selon leur score total et les items selon leur difficulté. Il en résulte une distribution plus ou moins en « escalier » illustrant le caractère hiérarchique des résultats.

Tableau 7 – Exemple d'items formant une échelle hiérarchique

Sujet #	Item 3	Item 2	Item 5	Item 1	Item 4	Total	# erreurs
3	1	1	1	1	1	5	0
10	1	1	1	1	0	4	0
9	1	1	1	1	0	4	0
4	1	1	*0*	*1*	0	3	*2*
5	1	1	1	0	0	3	0
2	1	1	1	0	0	3	0
7	1	1	*0*	*1*	0	3	*2*
6	1	0	0	0	0	1	0
1	1	0	0	0	0	1	0
8	0	0	0	0	0	0	0
Moyenne	0,9	0,7	0,5	0,5	0,1	2,7	Total = 4

Comme on peut le constater, les résultats aux cinq items de ce test forment une hiérarchie presque parfaite. L'item 4 est le plus difficile et l'item 3 est le plus facile. Lorsqu'un item difficile est réussi, tous les items plus faciles le sont aussi, mises à part quelques exceptions qui constituent des *erreurs* (sujets # 4 et 7). Règle générale, on peut donc affirmer que les résultats à ce test sont *reproductibles*. En effet, lorsqu'un test est hiérarchique, il est possible de prédire quels items ont été réussis et quels items ont été échoués à partir de la seule connaissance du score total. Connaissant l'ordre de difficulté des items et sachant que cet ordre est le même pour tous, un élève qui obtiendrait un score de 3 à l'examen du tableau 7, devrait réussir les items 3, 2 et 5. Lorsqu'un élève ayant obtenu un score de 3 réussit un autre item que les trois mentionnés précédemment, il y a erreur dans la reproductibilité totale du test : c'est le cas des sujets 4 et 7.

Guttman (1950) propose de considérer que les résultats d'un test sont hiérarchiques lorsque moins de 10% des résultats ne sont pas reproductibles. Il propose de calculer un coefficient de reproductibilité de la manière suivante :

$$CR = 1 - \frac{n_e}{n_j n_p} \qquad (5.10)$$

Dans cette équation, *CR* est le coefficient de reproductibilité, n_e le nombre d'erreurs de reproductibilité, n_j le nombre d'items et n_p le nombre de personnes. Le nombre d'erreurs est donné par le nombre de fois qu'un item fournit un résultat qui n'est pas en accord avec selon le score total obtenu et l'ordre de difficulté de l'ensemble des items.

Le coefficient de reproductibilité a plusieurs fois été employé comme preuve de la validité conceptuelle des tests opératoires piagétiens. Dans le cas des données du tableau 7, sa valeur est la suivante :

$$CR = 1 - \frac{4}{5 \times 10} = 0{,}92$$

Le test du tableau 7 possède donc une reproductibilité acceptable, supérieure au seuil de 0,90 recommandé par Guttman (1950). Sa validité conceptuelle, du point de vue de l'invariance de l'ordre de réussite de ses items a été démontrée.

Pour vérifier la validité conceptuelle d'une épreuve, les méthodes corrélationnelles sont également fort utiles. Les méthodes s'appuyant sur les corrélations sont de trois types :

1. l'évaluation des corrélations simples ;
2. les matrices multi-trait multi-méthode ;
3. l'étude des traits latents.

4.2 Validité conceptuelle et corrélation simple

La corrélation simple nous permet de déterminer si un test mesure la même chose qu'un autre test (validité convergente) ou encore quelque chose de complètement différent (validité divergente). Un test de créativité qui aurait une corrélation élevée avec un test d'intelligence verbale pourrait nous laisser perplexe sur ce que le test mesure réellement. Faudrait-il remettre en question le fait que le test mesure la créativité ou émettre l'hypothèse que le test mesure quelque chose de particulier, telle que la créativité verbale ?

Lorsqu'un test corrèle bien avec d'autres tests qui mesurent des variables voisines (p.e. intelligence et rendement scolaire) et qu'il ne corrèle pas avec des variables éloignées (p.e. intelligence et motivation), alors nous pouvons cerner de manière plus adéquate l'étendue des variables mesurées par notre instrument de mesure.

Lorsque l'on a voulu remplacer le test d'indépendance du champ de Witkins (*« Rod and Frame Test »*) par un test papier-crayon intitulé *« Test collectif des figures cachées » (TCFC)*, on s'est empressé de mesurer la corrélation entre les résultats aux deux tests. Comme il y avait effectivement une assez bonne corrélation entre les deux tests, tout semblait appuyer l'hypothèse que les deux tests mesuraient la même chose. Cependant, le *TCFC* n'a pas démontré la même validité divergente que le *« Rod and Frame Test »* avec les sous-tests d'intelligence non verbale, tel que le sous-test des blocs de Kohs-Goldstein. Le *TCFC* mesure donc l'intelligence non verbale en plus du style cognitif de dépendance/indépendance du champ. Le test papier-crayon ne mesure pas une habileté aussi épurée que celle mesurée par le test individuel originel de Witkins.

4.3 MATRICE MULTI-TRAIT MULTI-MÉTHODE

Campbell et Fiske (1959) ont proposé une approche rigoureuse à l'étude des validités convergentes et divergentes. Il s'agit de construire une matrice de corrélations entre résultats à des tests différents par ce qu'ils mesurent (*multi-trait)* et par la façon dont ils le mesurent (*multi-méthode).* Selon cette approche, la corrélation la plus forte devrait être obtenue entre deux tests mesurant le même trait avec la même méthode : c'est la fiabilité d'équivalence. La corrélation entre deux tests mesurant le même trait par des méthodes différentes (mono-trait, multi-méthode) nous fournit les validités convergentes. Les corrélations entre deux tests mesurant des traits différents par la même méthode (multi-trait, mono-méthode) ou par des méthodes différentes (multi-trait, multi-méthode) devraient nous fournir la validité discriminante. Ces deux derniers indices de validité devraient être significativement plus faibles que la fiabilité ou la validité convergente.

Le tableau 8 présente une matrice multi-trait multi-méthode. On y trouve les résultats d'une étude fictive de validité conceptuelle sur trois examens de mathématiques (calcul, géométrie, problèmes écrits) mesurés selon deux méthodes différentes (questions à choix multiple et questions à réponses courtes). En diagonale, nous retrouvons la corrélation de chaque test avec un test similaire employant la même méthode : il s'agit de la fiabilité d'équivalence (en gras). Celle-ci est très satisfaisante, mais il semble que les questions à choix de réponses donnent lieu à des résultats plus fiables (0,91 à 0,95) que les réponses courtes (0,82 à 0,85). Faut-il y voir un effet d'erreurs dues à la subjectivité dans la correction des réponses courtes ?

Tableau 8 – Exemple de matrice multi-trait multi-méthode

	Méthode 1 (Choix multiple)			Méthode 2 (Réponses courtes)		
	Calcul	Géométrie	Prob. écrits	Calcul	Géométrie	Prob. écrits
1. Choix multiple						
• Calcul	**0,95**					
• Géométrie	0,51	**0,92**				
• Problèmes écrits	0,42	0,26	**0,91**			
2. Réponses courtes						
• Calcul	*0,83*			**0,85**		
• Géométrie	0,28	*0,86*		0,41	**0,88**	
• Problèmes écrits	0,17	0,15	*0,79*	0,36	0,22	**0,82**

La validité convergente fournit des résultats satisfaisants : elle est donnée par la corrélation entre deux tests qui mesurent le même trait par des méthodes différentes. En effet, on est en droit de s'attendre à ce que des performances mathématiques semblables, mesurées par des méthodes différentes, possèdent une certaine corrélation

entre elles. Les corrélations sont très élevées (en italiques dans le tableau 8) : elles varient de 0,79 à 0,86. On peut interpréter ces résultats de deux manières. La première est que la méthode de mesure employée a peu d'effet sur la réussite et que le trait mesuré est vraiment la variable la plus importante. La seconde manière d'interpréter ce résultat est typiquement édumétrique : ces résultats pourraient également signifier que l'apprentissage est suffisamment généralisé pour permettre aux élèves de réussir des problèmes présentés différemment.

Les corrélations sous la diagonale de chacune des parties de la matrice multi-trait multi-méthode fournissent les coefficients de validité discriminante. On en distingue deux sortes : les corrélations mono-traits hétéro-méthodes et les corrélations hétéro-traits hétéro-méthodes. Ces derniers coefficients sont les plus faibles de tous (0,15 à 0,28). En effet, ces coefficients de validité discriminante font intervenir non seulement des traits mais aussi des méthodes de mesure différentes. Lorsque la validité discriminante ne porte que sur l'effet de la méthode de mesure, les corrélations vont de faibles à modérées (méthode 1 : 0,26 à 0,51 ; méthode 2 : 0,22 à 0,41). Il est à noter que les valeurs de validité discriminante pour la méthode 2 sont toutes inférieures à celles obtenues par la méthode 1. Ceci est dû au fait que les tests sont plus fiables avec la méthode 1, ce qui permet de meilleures corrélations entre les scores observés.

La matrice multi-trait multi-méthode nous révèle que, parmi les trois tests, ce sont *« Problèmes écrits »* et *« Géométrie »* qui mesurent les habiletés les plus indépendantes l'une de l'autre. Ce résultat pourrait s'expliquer par le fait que le test de problèmes écrits mesure aussi des habiletés de compréhension de lecture qui ne sont pas requises pour mesurer les habiletés de géométrie. Si cette observation devait confirmer les hypothèses d'une quelconque théorie didactique de l'apprentissage des mathématiques, ceci contribuerait à accroître notre confiance en la validité conceptuelle de ces deux instruments de mesure.

4.4 ÉTUDE DES TRAITS LATENTS

Dans plusieurs situations, nous savons bien qu'en dépit des différences de contenu, de format, de tâches des items, ceux-ci mesurent une caractéristique commune, qui les influence tous. Nous nous attendons dans ce cas à ce que les items qui mesurent une même caractéristique soient fortement corrélés. Comme pour l'analyse des résultats aux tests par la matrice multi-trait, multi-méthode, des items (ou des sous-tests) mesurant le même trait devraient se réunir en « grappes » de corrélations élevées, que l'inspection visuelle d'une matrice de corrélations devrait nous révéler.

L'analyse factorielle permet d'aller plus loin que la simple inspection visuelle des matrices de corrélation. Elle permet également d'extraire les composantes ou facteurs de variance commune, chaque facteur rendant compte d'une partie de la variance totale des résultats qui n'est pas expliquée par les autres facteurs. Ces facteurs ou composantes principales sont également appelés *traits latents,* parce qu'il s'agit de variables non observées et construites sur lesquelles on projette la variance commune à un certain nombre de variables mesurant la même caractéristique.

La figure 2 illustre de façon simple ce qu'est un trait latent. La situation présentée est celle de deux variables A et B en forte corrélation. Si A et B sont fortement corrélées, c'est que, vraisemblablement, elles mesurent la même chose, le même trait

latent. On peut se demander pourquoi il est nécessaire d'utiliser deux variables pour mesurer la même chose, alors qu'elles nous fournissent toutes deux la même information. Quelle variable faut-il conserver ?

La solution est représentée graphiquement dans la figure 2. Plutôt que d'effectuer les observations dans un système à deux variables, celles-ci peuvent être projetées sur une nouvelle variable qui retient l'essentiel de la variance commune à A et à B. Cette nouvelle variable est le *trait latent*, généralement appelé *facteur*. Comme on peut le constater, cette opération, qui réduit le système d'observation de deux à une seule variable, n'entraîne qu'une légère perte d'information puisque toutes les observations ne coïncident pas parfaitement avec la droite représentant le facteur. Toutefois, ce dernier rend compte de la plus grande partie de la dispersion des résultats qui s'effectue à présent selon l'axe horizontal. Quant à la dispersion des résultats selon l'axe vertical, elle correspond à une quantité négligeable. Ce qui est légèrement perdu en information est largement gagné en parcimonie, c'est-à-dire en simplicité du modèle.

La figure 2 présente une situation d'unidimensionnalité, c'est-à-dire qu'un seul facteur est nécessaire pour rendre compte de la variance des résultats. Lorsqu'il faut plus d'un facteur pour expliquer les résultats, nous avons affaire à un modèle *multidimentionnel* de traits latents (figure 3).

Cette situation se présente lorsque A et B ne sont que modérément corrélés. Il est alors difficile d'expliquer la variance commune entre ces deux variables par un système à une seule variable. Une fois expliquée une bonne partie de la dispersion des résultats par un axe horizontal représentant le trait latent, il subsiste une forte dispersion selon l'axe vertical dont nous ne rendons pas compte. Nous sommes alors face à un choix :

1. ne retenir qu'une seule dimension avec le risque de ne pas rendre compte d'une partie importante de la variance des résultats (représentée par la dispersion verticale) ;
2. ajouter une deuxième dimension avec la perte de parcimonie que cela implique.

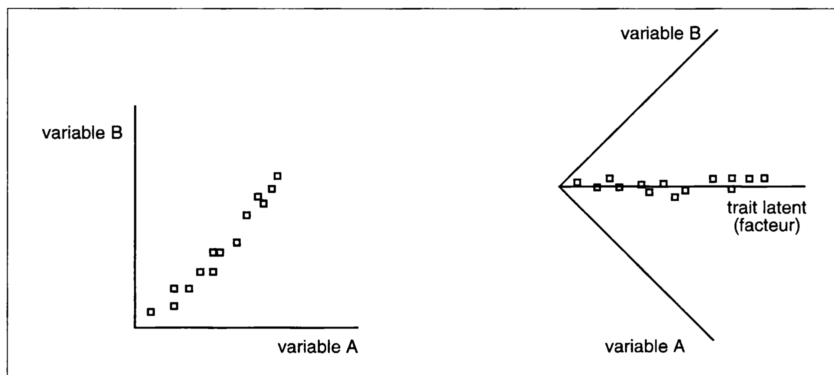


Figure 2 – Le trait latent explique la plus grande partie de la variance

Ce choix doit se faire en pondérant les avantages qu'il y a à remplacer deux variables par un seul trait latent et les inconvénients qu'il y a à laisser tomber une partie de la variance totale. Lorsque cette part est trop grande, il faut utiliser plusieurs

traits latents, du moins tous ceux qui sont nécessaires pour expliquer une part substantielle de la variance des résultats.

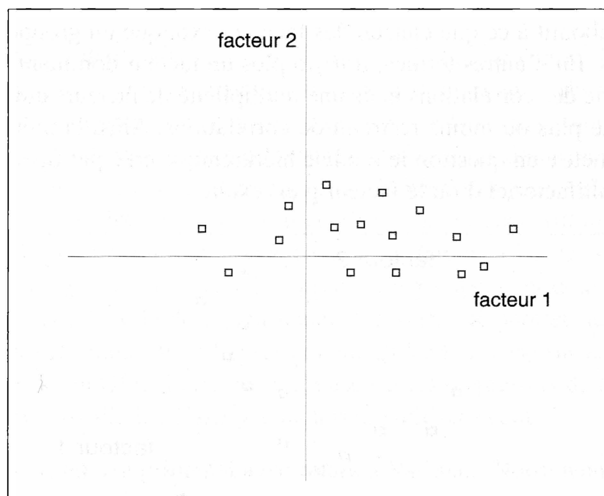


Figure 3 – Modèle avec deux facteurs

Lorsqu'une analyse factorielle doit porter sur les *j* items à un examen, nous sommes loin de la situation relativement simple décrite par un système à deux variables. C'est de *j* variables qu'il s'agit et donc, potentiellement de *j* traits latents. Idéalement, le constructeur de test préférera se trouver dans une situation où ces traits latents sont peu nombreux et faciles à discerner. Le cas le plus simple est celui d'un test unidimensionnel ne comprenant qu'un seul trait latent. Par contre, lorsqu'un seul trait latent n'est pas suffisant pour expliquer la plus grande partie des résultats, il faut avoir recours à d'autres traits latents. S'il y a trop de traits latents, c'est que le test mesure une grande variété de caractéristiques # : à la limite, presqu'autant de caractéristiques différentes qu'il y a d'items. Pour des raisons de simplicité d'interprétation, le constructeur de test préfère se retrouver dans la situation où son test ne mesure pas trop de variables sans rapport entre elles.

Le développement de l'analyse factorielle est intimement lié à l'histoire des tests. C'est en effet Spearman (1907) qui, au début du siècle, jette les bases de l'analyse factorielle. Observant des corrélations élevées entre les résultats à différents tests d'intelligence, Spearman avance l'hypothèse que les performances à ces tests sont essentiellement déterminées par un facteur général, le *facteur g*. Des facteurs spécifiques à chaque test interviennent également mais jouent un rôle mineur. La méthode d'analyse factorielle développée par Spearman lui permet de produire des résultats empiriques en faveur de son hypothèse. Trente ans plus tard, ces résultats sont toutefois remis en question par Thurstone qui s'appuie sur une nouvelle technique d'analyse factorielle. Comme Spearman, Thurstone (1928) utilise des axes factoriels orthogonaux et donc indépendants les uns des autres. Cependant, plutôt que de maintenir ces axes de façon à ce que le premier facteur explique la plus grande partie de la variance et que les autres n'en expliquent que le résidu, il a l'idée d'effectuer une rotation des axes afin d'améliorer le degré d'adaptation entre les données et la structure factorielle.

Il recherche ainsi la structure la plus simple et détermine celle-ci par des critères mathématiques dont le plus connu est certainement le critère *Varimax*, suivant lequel on cherche à ce que la variance soit maximum sur chacun des axes factoriels (figure 4). Cette méthode aboutit à ce que chacun des facteurs explique un groupe de résultats et rien que celui-là. En d'autres termes, il n'y a plus un facteur dominant qui explique la plus grande partie des corrélations mais une multiplicité de facteurs qui, chacun, explique un ensemble plus ou moins restreint de corrélations. Ainsi la méthode de Thurstone conduit à mettre en question le modèle hiérarchique créé par Spearman, au profit d'un modèle multifactoriel d'où le facteur g est exclu.

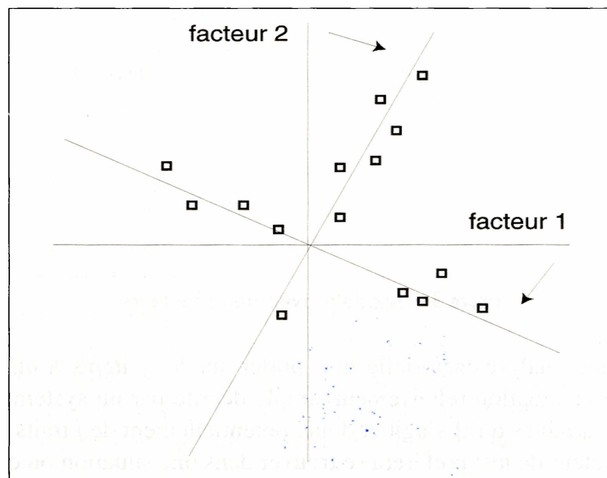


Figure 4 – Exemple d'application de la rotation Varimax

Le débat entre Spearman et Thurstone illustre bien l'intérêt de l'analyse factorielle comme moyen de validation d'un modèle de traits latents. Mais il souligne aussi les limites de cette méthode. En effet, du point de vue strictement mathématique, Spearman et Thurstone ont tous les deux raisons. En fait, l'analyse factorielle permet seulement de vérifier si les données sont consistantes ou non avec la structure factorielle postulée. Lorsque les données sont compatibles avec plusieurs structures latentes, l'analyse factorielle ne nous permet pas de déterminer laquelle choisir. Ce choix ne peut être fait que sur des bases théoriques. Par conséquent, dans le cadre d'une démarche de validation d'un test, l'analyse factorielle nous apporte des informations nécessaires mais non suffisantes. Le praticien ne devra donc pas oublier que toute démarche d'analyse factorielle s'appuie sur deux postulats de base :

1. *Le postulat de causalité factorielle* selon lequel les variables observées sont des combinaisons linéaires de variables causales sous-jacentes. Les résultats d'analyse factorielle ne peuvent, en aucun cas, nous servir à prouver ce postulat. Ces résultats peuvent éventuellement nous amener à conclure qu'un modèle factoriel, basé sur ce postulat, n'est pas consistant avec les données d'observation. Par contre, lorsqu'il y a consistance, il reste encore au chercheur à défendre la pertinence du modèle qu'il propose.

2. *Le postulat de parcimonie* selon lequel, entre deux solutions factorielles, nous devrons choisir la plus simple. Bien que ce postulat soit largement accepté par

les chercheurs, il n'est pas possible de démontrer son bien-fondé. Dans la réalité, une structure factorielle simple est-elle toujours plus plausible qu'une structure plus complexe ?

Du fait de ces deux postulats, l'utilisation de l'analyse factorielle comme technique de validation est moins évidente qu'il n'y paraît au premier abord. Pour illustrer la complexité de l'interprétation des résultats d'analyse factorielle, nous avons soumis les mêmes données à deux analyses différentes. Il s'agit des données d'étalonnage de l'adaptation française de l'échelle d'intelligence de Wechsler pour enfants, le WISC-R (Wechsler, 1981). Ces données ont été recueillies sur un échantillon de 1066 sujets représentatif de la population française âgée de 6 ans 6 mois à 16 ans 6 mois. Pour rappel, ce test d'intelligence comprend 12 épreuves regroupées en deux ensembles # : l'échelle Verbale et l'échelle de Performance. Le WISC-R permet de calculer un QI Total, basé sur les résultats aux 12 épreuves, un QI Verbal, basé sur les 6 épreuves de l'échelle Verbale, et un QI de Performance, basé sur les 6 épreuves de l'échelle de Performance. Les trois méthodes d'analyse factorielle utilisées sont :

(1) Une analyse en axe principal avec rotation Varimax. Nous avons défini a priori une solution avec deux facteurs.

(2) Une analyse en axe principal avec rotation Varimax pour une solution à trois facteurs spécifiée a priori. De nombreux chercheurs (par exemple, Kaufman, 1975) ont en effet défendu l'idée d'un regroupement des épreuves en trois ensembles, au lieu de deux. À côté d'un facteur « Compréhension Verbale » et d'un facteur « Organisation Perceptive », ces auteurs postulent l'existence d'un facteur « Attention/Concentration » qui saturerait particulièrement les épreuves *Mémoire*, *Code* et *Arithmétique*.

Tableau 9 – Analyse factorielle en axe principal avec rotation Varimax (solution avec deux facteurs)

Epreuves	Facteur 1	Facteur 2
Vocabulaire	**0,83**	0,19
Compréhension	**0,73**	0,20
Information	**0,73**	0,29
Similitudes	**0,69**	0,32
Arithmétique	**0,58**	0,30
Mémoire	**0,46**	0,18
Ass. d'Objets	0,18	**0,74**
Cubes	0,27	**0,67**
Compl. d'Images	0,36	**0,57**
Arrang. d'Images	0,37	**0,51**
Labyrinthes	0,12	**0,44**
Code	0,23	0,22

Tableau 10 – Analyse factorielle en axe principal avec rotation Varimax (solution avec trois facteurs)

Epreuves	Facteur 1	Facteur 2	Facteur 3
Vocabulaire	**0,80**	0,19	0,27
Compréhension	**0,72**	0,20	0,23
Information	**0,68**	0,28	0,27
Similitudes	**0,61**	0,30	0,31
Ass. d'Objets	0,16	**0,74**	0,12
Cubes	0,14	**0,65**	0,35
Compl. d'Images	0,36	**0,58**	0,12
Arrang. d'Images	0,37	**0,52**	0,11
Labyrinthes	0,09	**0,43**	0,17
Arithmétique	0,40	0,24	**0,55**
Mémoire	0,27	0,10	**0,53**
Code	0,13	0,17	**0,42**

Pour comprendre correctement les données figurant dans les tableaux 9 et 10, quelques explications techniques sont nécessaires. Les valeurs mentionnées dans ces tableaux représentent les *saturations* des épreuves par chacun des facteurs. Lorsque les différents facteurs sont orthogonaux, c'est-à-dire non corrélés (les axes factoriels forment alors un angle de 90°), les saturations sont les corrélations entre les facteurs et les variables. C'est le cas dans nos deux exemples. Par conséquent, en élevant une saturation au carré, nous obtenons la proportion de variance d'une variable déterminée par le facteur en question. Passons à présent en revue les tableaux. Nous pouvons nous rendre compte que les deux analyses factorielles réalisées à partir des mêmes données apportent des arguments en faveur de deux modèles factoriels possibles. La solution avec deux facteurs va dans le sens du regroupement des épreuves du WISC-R en deux sous-échelles, l'une Verbale et l'autre de Performance. Dans ce modèle, seule l'épreuve de Code ne montre pas de saturations factorielles bien affirmées. La solution avec trois facteurs rend admissible un autre regroupement d'épreuves. Quelle solution factorielle devons-nous dès lors choisir ? Comme souligné plus haut, la réponse n'est pas de nature mathématique. C'est en fait le modèle du fonctionnement intellectuel que nous défendons qui permettra de déterminer la solution factorielle la plus adéquate.

Hormis les problèmes d'interprétation, l'analyse factorielle soulève plusieurs questions méthodologiques relatives aux conditions de son application. Les plus importantes concernent :

(1) *La taille de l'échantillon.* Plus l'échantillon de sujets est petit, moins les coefficients de corrélation entre les variables observées seront fiables. Par conséquent, les solutions factorielles obtenues seront sujettes à caution. Il n'y a toutefois pas de taille d'échantillon idéale. Une règle généralement admise est d'avoir au moins cinq sujets par variable observée (Gorsuch, 1983). Par exem-

ple, si nous souhaitons réaliser une analyse factorielle avec les réponses à un questionnaire de 40 questions, celui-ci devra être rempli par au moins 200 sujets. Cette règle n'est cependant pas absolue. Si les corrélations entre variables sont très élevées et très fiables et que les facteurs sont peu nombreux, un échantillon relativement petit pourra suffire. Par contre, si les corrélations entre variables sont toutes faibles (inférieures à 0,30), l'opportunité de réaliser une analyse factorielle devra être remise en question, quelle que soit la taille de l'échantillon. En effet, dans un tel cas, il n'y a pratiquement rien à analyser. Par conséquent, avant de réaliser une analyse factorielle, une inspection de la matrice des corrélations entre variables s'impose.

(2) *La normalité.* Les inférences statistiques utilisées pour déterminer le nombre de facteurs s'appuient sur le postulat de normalité multivariée. Ce postulat signifie que toutes les variables et toutes les combinaisons de variables se distribuent normalement. Nous ne pouvons tester la normalité de toutes les combinaisons linéaires de variables. Par contre, la normalité de chaque variable peut être appréciée en regardant son coefficient d'asymétrie et son coefficient d'aplatissement (voir chapitre 1).

(3) *La linéarité.* Rappelons que les coefficients de corrélation évaluent une relation linéaire entre les variables. En cas de non linéarité de ces relations, les coefficients de corrélations en seront affectés, ce qui risque de remettre en question les résultats des analyses factorielles. La linéarité de la relation entre variables peut être vérifiée à l'aide de graphiques en nuage de points.

5. La validité différentielle

5.1 Le concept de biais

La validité d'un test est généralement évaluée pour l'ensemble de la population pour laquelle le test a été développé. On postule ainsi que la validité d'une inférence faite à partir des scores au test en question est toujours équivalente pour tous les sujets de cette population. Depuis les années 70, ce postulat a été largement remis en question. Nous ne pouvons en effet pas écarter a priori que la validité d'un test puisse varier au sein d'une même population selon le groupe d'appartenance des sujets évalués. Par exemple, un test de mathématique peut nous permettre d'évaluer de manière valide les compétences en résolution de problème à condition que les sujets n'aient aucune difficulté pour lire les énoncés des questions. Par conséquent, ce test ne sera pas valide pour évaluer des sujets souffrant de troubles de la lecture. Ces sujets obtiendront systématiquement des scores faibles du fait de leur difficulté à déchiffrer les questions et non du fait de leur niveau de compétence en résolution de problèmes mathématiques. La validité du test variera donc au sein d'une même population selon que le sujet évalué appartienne ou non au groupe des mauvais lecteurs. De même un test intellectuel peut avoir une validité différente pour les filles et pour les garçons s'il est constitué uniquement de problèmes de nature spatiale. En effet, les filles ont habituellement plus de difficultés que les garçons à réaliser des opérations sur des représentations spatiales. Elles risquent dès lors d'avoir des scores systématiquement inférieurs à ceux des garçons alors que leur capacité de raisonnement est identique.

Il apparaît donc nécessaire d'évaluer la validité d'un test non seulement pour les différents usages que nous souhaitons en faire mais aussi pour les différents groupes de la population auxquels nous aurons l'occasion de l'appliquer. On parle à ce propos de l'étude de la validité différentielle du test. Un biais existe lorsqu'une différence de validité du test est observée entre certains groupes de la population. En d'autres termes, un test est biaisé lorsque « *les scores à ce test ont des significations ou des implications pour un groupe déterminé d'utilisateurs qui diffèrent de leurs significations ou de leurs implications pour les autres utilisateurs* » (Cole & Moss, 1989, p. 205). L'évaluation de la *validité différentielle* d'un test est une procédure complexe. Comme toute étude de validité, il s'agit d'une démarche toujours inachevée. Pour chacune des utilisations escomptées du test, il est en effet nécessaire de produire des preuves de l'absence de biais. L'évaluation différentielle de la validité de contenu, de la validité critérielle et de la validité conceptuelle apporte trois types de preuves complémentaires concernant l'absence de biais dans le test étudié.

Soulignons d'emblée que l'existence d'une différence de moyenne entre les scores de deux groupes de la population n'est pas en soi la preuve de l'existence d'un biais. En fait, une différence de moyenne peut simplement refléter une différence d'opportunité d'apprentissage entre les deux groupes de sujets considérés. Par exemple, si les filles choisissent moins souvent que les garçons les options scientifiques, il sera logique d'observer à un test de sciences un score moyen des filles inférieur à celui des garçons. Dans ce cas, nous ne pourrons bien entendu pas parler de biais. Dans les tests cognitifs et d'acquis scolaires, l'observation de différences d'efficience est inévitable car celles-ci reflètent les différences d'opportunités d'apprentissage offertes à chacun par son milieu. Par conséquent, « *c'est l'absence de différence observée qui devrait poser problème et mettre en doute la qualité d'un test, et non l'inverse* » (Grégoire, 1992, p.93). Les tests, en permettant de mettre en évidence les différences de performances entre les groupes qui composent la population, peuvent d'ailleurs avoir une utilité sociale. Grâce à de telles observations, nous sommes conduits à mettre en oeuvre des actions de remédiation dont l'objectif est de donner à chacun des chances d'épanouissement et de réussite les plus égales possibles.

5.2 ÉVALUATION DE LA VALIDITÉ DIFFÉRENTIELLE

Nous avons indiqué plus haut que, pour repérer les éventuels biais dans un test, nous devons vérifier que la validité de ce test est équivalente pour les différents groupes de la population. Pour ce faire, nous devons examiner les différentes facettes de la validité du test : son contenu, les prédictions qu'il nous permet de réaliser, les scores qu'il nous permet de calculer, les différences qu'il nous permet d'évaluer entre les sujets... Traditionnellement, les différentes facettes de la validité sont rassemblées dans les trois catégories que sont la validité de contenu, la validité critérielle et la validité conceptuelle. Nous allons tour à tour les examiner.

5.2.1 *La validité de contenu*

L'évaluation de la validité différentielle du contenu consiste à vérifier si, au sein de chacun des groupes de la population, le contenu des items est approprié pour mesurer la réalité souhaitée. Cette évaluation s'appuie sur les jugements de spécialistes du domaine mesuré par le test. Ces jugements concernent les représentations et la familia-

rité des membres de chaque groupe par rapport au contenu des items. Ils concernent également la présence de stéréotypes relatifs à l'un des groupes en question qui pourraient éventuellement favoriser ou défavoriser les performances. Malheureusement, ces jugements ont l'inconvénient de rester souvent très subjectifs. On se contente généralement de passer en revue tous les items et d'éliminer ceux qui paraissent inadaptés pour certains groupes. La limite de cette méthode est bien exprimée en anglais par l'expression ironique qui la désigne # : « *armchair validity* ». Pour diminuer la subjectivité des jugements, des grilles d'analyse ont été mises au point et l'évaluation des items est généralement faite par plusieurs juges. La reconnaissance d'un item comme biaisé est alors décidée sur base de l'ensemble des jugements. Mais les résultats ne semblent pas à la hauteur de l'effort fourni car les tests ainsi « nettoyés » des items biaisés ne donnent le plus souvent pas des résultats très différents de ceux obtenus avec les tests originels (Flaugher, 1978 ; Sattler, 1988). La détection des biais ne peut donc se limiter à la seule évaluation par des juges. Cette méthode doit être complétée par une évaluation quantitative basée sur les résultats obtenus par les différents groupes étudiés.

Les évaluations quantitatives s'intéressent essentiellement à la difficulté et à la discrimination des items. Elles ont pour but de vérifier si tous les items permettent de classer les sujets de manière équitable. Pour cela, les items doivent mesurer uniquement la réalité que nous désirons évaluer et non des variables parasites liées au groupe d'appartenance. Si, par exemple, dans un test de raisonnement, certains items font appel aux règles d'un sport très pratiqué par les garçons mais peu par les filles, ces items risquent d'être inéquitables. Ils seront en effet plus faciles pour les garçons que pour les filles du fait de l'influence d'une variable qui n'a rien à voir avec les capacités de raisonnement. De tels items présentent un fonctionnement différentiel qui conduit habituellement à les éliminer du test. Le fonctionnement différentiel d'un item n'est pas uniquement lié au contenu de la question. Il peut aussi découler des modalités de réponse à cet item. Par exemple, pour certains groupes de sujets, le système de réponse à choix multiple peut être une source de difficulté particulière. De même, si certains items demandent une réponse écrite, la qualité de la calligraphie des sujets peut être source d'iniquité. Certains correcteurs peuvent en effet être influencés favorablement ou défavorablement dans leur cotation par la calligraphie du texte dont ils doivent juger le contenu. Comme nous pouvons le voir, les sources d'iniquité sont nombreuses et demandent une analyse minutieuse du fonctionnement différentiel de tous les items du test dont nous évaluons la validité différentielle.

L'analyse du fonctionnement différentiel des items se fait généralement lors de la construction du test. Pour cette raison, nous détaillons les techniques d'analyse du fonctionnement différentiel dans le chapitre consacré à l'analyse des items (chapitre 6, section 6). Par ailleurs, des techniques plus sophistiquées d'analyse du fonctionnement différentiel, basées sur les modèles de la réponse à l'item, sont décrites dans le chapitre consacré à la présentation de ces modèles (chapitre 8, section 5).

5.2.2 *La validité en référence à un critère*

Lorsque nous étudions la validité différentielle d'un test, il est souvent important de comparer, pour différents groupes de la population, la relation entre le score au test et une mesure externe prise comme critère. En effet, cette relation sous-tend de

nombreuses décisions prises à partir des résultats de tests. Par exemple, des enfants sont régulièrement orientés dans l'enseignement spécialisé sur base de leurs faibles résultats à un test intellectuel dont les scores sont très liés à la réussite scolaire. De même, des étudiants peuvent se voir refuser l'accès à un programme d'étude du fait de leur score insuffisant à un test prédictif de la réussite de ce programme. Vu l'importance de ces décisions, il est essentiel que la validité prédictive d'un test soit équivalente pour tous les groupes concernés par l'usage de ce test. Pour contrôler si un test est équitable du point de vue prédictif pour deux groupes de sujets, nous pouvons calculer dans chacun de ces deux groupes les coefficients de corrélation entre les résultats au test et les résultats au critère et vérifier s'il existe une différence significative entre ces coefficients. La comparaison des coefficients de corrélation est cependant insuffisante. À deux coefficients identiques peuvent en fait correspondre des systèmes de prédiction différents. Pour nous en rendre compte, nous devons déterminer, pour les deux groupes étudiés, la droite de régression qui unit les scores au test et au critère. Si cette droite est identique dans les deux groupes, nous pouvons, en première approximation, considérer que la validité prédictive du test est équitable pour les deux groupes considérés. Si les droites de régression sont au contraire différentes pour chaque groupe, le test doit être considéré comme biaisé car il conduit à des prédictions différentes en fonction du groupe d'appartenance.

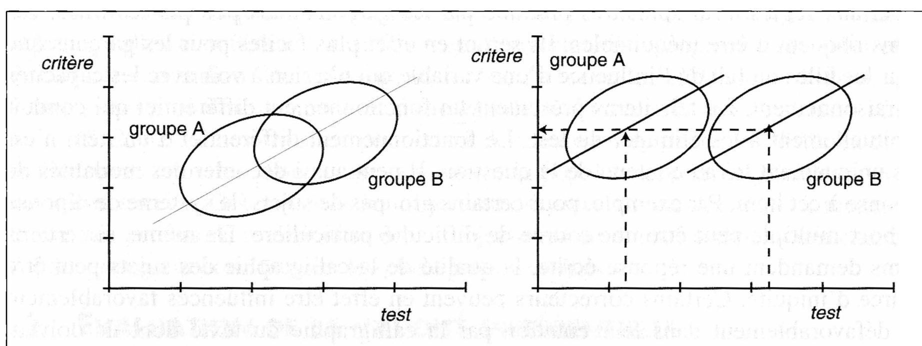


Figure 5 – Évaluation de la validité prédictive différentielle à l'aide des droites de régression des deux groupes comparés

La figure 5 propose une illustration de ces deux situations. Les ovales représentent les nuages de points pour chacun des deux groupes. Dans le graphique de gauche, bien que le score moyen au test soit différent dans les deux groupes, nous constatons que la droite de régression est la même pour les deux groupes. Par conséquent, quel que soit le groupe d'appartenance, un score élevé au test implique un résultat élevé au critère, et réciproquement. Dans le graphique de droite, le score moyen au test est différent dans les deux groupes mais les droites de régression sont également très différentes. Par conséquent, les prédictions faites sur base des scores au test sont biaisées. Si un sujet appartient au groupe B, il devra en effet obtenir un score beaucoup élevé au test qu'un sujet du groupe A pour que la prédiction du résultat au critère soit la même (droites fléchées en pointillés).

La comparaison des droites de régression soulève toutefois quelques problèmes d'interprétation. Les erreurs de mesure au test et au critère peuvent en effet être différentes selon les groupes. Par conséquent, du seul fait de l'inégalité des erreurs de mesure, des différences de droite de régression peuvent apparaître entre certains groupes alors que le test n'est pas biaisé. Lors de l'évaluation de la prédiction différentielle, nous devrons donc toujours tenir compte des erreurs de mesure dans chacun des groupes considérés. Par ailleurs, l'importance de la pente de la droite peut également entraîner un biais en défaveur d'un des groupes. Cette situation est illustrée dans la figure 6 (d'après Camilli & Shepard, 1994) où les deux groupes partagent la même droite de régression. Toutefois, comme nous pouvons le constater, la différence de moyenne est plus grande sur le test que sur le critère.

Si la distribution des scores est normale dans les deux groupes au test et au critère, nous pouvons aisément nous rendre compte que cette situation aboutit à une injustice à l'encontre des sujets du groupe A. Supposons que, pour sélectionner les sujets, nous fixions le score seuil au niveau du score moyen du groupe B. Dans ce cas, 50% des sujets de ce groupe seront sélectionnés. Par contre, dans le groupe A, 16% seulement des sujets seront sélectionnés. Pourtant, sur le critère, 31% des sujets de ce même groupe atteignent le niveau de performance désiré. Pour résoudre ce problème, Thorndike (1971) propose de choisir un score seuil au test différent pour les deux groupes en fonction de leur performance sur le critère.

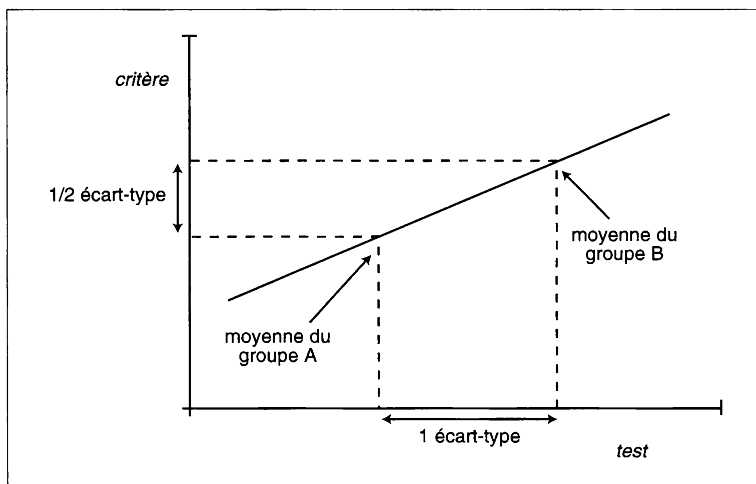


Figure 6 – Présence d'un biais malgré une droite de régression commune aux deux groupes comparés

5.2.3 *La validité conceptuelle*

Lorsque nous apprécions la validité différentielle d'un test, nous devons nous demander si l'organisation du test, basée sur un modèle théorique donné, est valide pour les différents groupes qui composent la population. Cette organisation sous-tend en effet le calcul des scores et des sous-scores. Il est donc essentiel de vérifier si les regroupements d'items et d'ensembles d'items sont fondés quel que soit le groupe d'appartenance des sujets évalués. Par exemple, le test d'intelligence de Wechsler pour

enfants (WISC-R) est organisé en deux échelles appelées respectivement « Verbale » et « Performance ». Pour que le calcul d'un score d'échelle ait un sens, il est nécessaire que les performances aux épreuves qui composent cette échelle soient très liées entre elles. Si ce n'est pas le cas, le score d'échelle ne sera guère plus qu'un amalgame d'informations hétérogènes sans signification précise. Pour légitimer l'organisation du WISC-R en deux échelles, l'analyse factorielle a été largement utilisée. Le plus souvent, les résultats de ces analyses ont confirmé le bien fondé de l'organisation du WISC-R pour les différents groupes de la population. Ainsi, Reschly (1978) a étudié la validité différentielle de l'organisation du WISC-R en fonction de l'origine ethnique des enfants américains # : blanche, noire, hispanique et amérindienne. À partir des résultats de chaque groupe, il a réalisé une analyse en composantes principales avec rotation varimax pour les solutions avec deux, trois et quatre facteurs. La solution avec deux facteurs apparaît comme la meilleure dans les quatre groupes étudiés. Cette solution bifactorielle recouvre la division du WISC-R en *Verbal* et *Performance*, et ceci pour tous les groupes. Ces résultats constituent un argument important en faveur de la validité différentielle du WISC-R.

Bien qu'apparemment simple dans son principe, l'usage de l'analyse factorielle pour étudier la validité différentielle d'un test soulève cependant une importante difficulté. Comment évaluer les similitudes et les différences entre les solutions factorielles obtenues dans différents groupes ? Il est en effet fréquent que les solutions obtenues ne soient pas tranchées. Il faut alors estimer si les solutions obtenues dans les différents groupes sont suffisamment proches pour être considérées comme équivalentes. Une comparaison purement subjective n'est pas suffisante et conduit à des conclusions peu consistantes. Des procédures quantitatives de comparaison ont été mises au point mais se révèlent relativement complexes à utiliser. C'est, par exemple, le cas de la procédure proposée par Jöreskog (1971), basée sur un *modèle structural d'équations*.

CHAPITRE 6

L'ANALYSE DES ITEMS

L'analyse des items ressemble à une répétition d'orchestre. Dans un orchestre, les instruments doivent jouer de façon harmonieuse. Selon la partition, certains interviendront à un moment bien précis. D'autres devront jouer en harmonie. Le tout doit produire une sensation musicale particulière correspondant aux intentions du compositeur et du chef d'orchestre.

Une situation similaire prévaut lors de l'analyse d'items. Celle-ci doit nous permettre d'identifier les items qui ne jouent pas en harmonie avec les autres ou qui ne jouent pas au même rythme. Certains jouent trop forts, d'autres pas assez. Certains se trompent carrément de partition. Le but du constructeur de test est de s'assurer que le message fourni par les items soit clair, harmonieux et précis. En psychométrie, l'analyse des items aide le constructeur de tests à choisir les meilleurs items à partir d'un ensemble de départ souvent plus grand que nécessaire. En éducation, la situation est toute autre. Les examens de rendement scolaire sont rarement mis à l'épreuve avant la passation en salle de classe. Ceci rend l'analyse d'items encore plus essentielle. C'est alors le seul moyen dont l'enseignant dispose pour modérer les résultats à un examen.

L'analyse d'items peut prendre plusieurs formes. Celles-ci dépendront des objectifs du constructeur de test et aussi de la méthode de préparation du test. En psychométrie, il est généralement prévu au départ de construire plus d'items que nécessaire, afin de ne retenir que ceux qui sont les plus valides. L'analyse des items correspond davantage à un processus de sélection : seuls les meilleurs seront retenus. En édumétrie, c'est la fonction de l'évaluation qui décide de l'analyse d'items. L'analyse d'items d'un examen final, administré en vue d'une évaluation sommative, sera fort différente de celle d'un instrument de mesure critériée, administré en vue d'une évaluation diagnostique ou d'une évaluation formative. Il se peut qu'un item convenant parfaitement dans le cadre d'une évaluation formative ne possède pas les caractéristiques désirées pour une bonne évaluation sommative.

Parmi les caractéristiques qui peuvent nous aider à mieux sélectionner les items en fonction des objectifs d'évaluation d'un test, les quatre suivantes sont les plus importantes :

- l'indice de difficulté ;
- l'indice de discrimination ;
- l'indice de fidélité ;
- l'indice de validité.

Malheureusement, il n'est pas possible d'interpréter ces indices en eux-mêmes. Chacun doit être interprété en fonction du contexte dans lequel l'instrument dont il fait partie est employé. Par exemple, il n'est pas possible d'affirmer qu'un item réussi par 90% des sujets est trop facile. La difficulté de l'item est relative au groupe (fort ou faible), mais aussi aux attentes face au groupe. S'agit-il d'un item mesurant un prérequis ? un objectif essentiel ? un objectif intermédiaire ? une aptitude complexe ? S'agit-il d'un groupe fort ? d'un groupe faible ? L'interprétation de la difficulté de l'item, ainsi que de toutes ses autres caractéristiques dépendra de la réponse que nous ferons à ces questions.

1. La difficulté de l'item

1.1 L'INDICE DE DIFFICULTÉ

La difficulté de l'item est donnée par la proportion des répondants qui réussit l'item — dans le cas d'items dichotomiques — ou encore par la moyenne des cotes accordées à cet item pour l'ensemble des sujets. L'expression suivante exprime la difficulté de l'item comme la somme de tous les résultats x obtenus à l'item, divisée par le nombre n de sujets :

$$p = \frac{\sum x}{n} \tag{6.1}$$

Plus la moyenne est élevée, plus l'item est réussi par un grand nombre de sujets. Plus elle est faible, moins l'item est réussi. Le tableau 1 présente les résultats du calcul de la moyenne pour trois items notés sur des échelles différentes. Le premier item est noté sur une échelle de cinq points, le second sur une échelle de deux points et le dernier sur une échelle dichotomique. Comme on le voit, les moyennes ne permettent pas de comparer la difficulté relative à chaque item. Par contre, comme nous l'avons vu au chapitre 4, la somme des moyennes des items nous permet de retrouver la moyenne des scores totaux au test. C'est ce que nous fournit la somme des résultats sur l'avant-dernière ligne : 3,1 + 1,4 + 0,6 = 5,1.

Lorsque nous désirons comparer la difficulté relative de plusieurs items, il nous faut ramener leurs moyennes à une échelle comparable. Il peut être compliqué d'analyser les résultats à un test dont les items sont notés sur des échelles différentes. Pour contourner ce problème, nous pouvons diviser la moyenne de chaque item par l'étendue de la note, ce qui produit une décimale (entre 0 et 1) que l'on peut interpréter de manière uniforme : c'est ce que nous appellerons l'*indice de difficulté*, afin de ne pas le confondre avec la *moyenne de l'item*. Dans le tableau 1 (dernière rangée), il ressort

clairement de cette transformation que c'est l'item corrigé sur deux points qui est le plus facile (p = 0,67).

La dernière colonne du tableau 1 nous fournit une autre valeur intéressante : celle de la difficulté moyenne des items (m_p = 0,63). Cette valeur est souvent préférable à la moyenne du test puisque cette dernière est influencée par le système de notation. En effet, il vaut mieux dire qu'un test a une difficulté moyenne de 0,63 plutôt que d'indiquer que la moyenne obtenue est 5,1 sur 8. L'indice de difficulté moyen ne tient pas compte de la pondération individuelle des items.

Tableau 1 – Moyennes et indices de difficulté de trois items

Sujet#	Item (/5)	Item (/2)	Item (/1)	Total (/8)
1	3	2	1	6
2	5	2	0	7
3	5	2	0	7
4	5	2	1	8
5	4	2	1	7
6	3	1	1	5
7	2	1	1	4
8	2	1	1	4
9	0	0	0	0
10	2	1	0	3
Moyenne	3,1	1,4	0,6	5,1
Difficulté P	0,62	0,67	0,56	0,63

Deux facteurs peuvent influencer notre interprétation de l'indice de difficulté :

- le nombre de réponses omises ;
- la probabilité de réussir l'item au hasard.

Lorsqu'un grand nombre de personnes n'ont pu répondre à un item par manque de temps, l'indice de difficulté ne reflète pas véritablement la difficulté de l'item. Plusieurs sujets n'ayant pas répondu auraient pu réussir un ou plusieurs items additionnels s'ils avaient disposé de plus de temps. Dans une telle situation, l'indice de difficulté mesure deux choses : la difficulté de l'item et la rapidité du répondant. Le calcul d'un nouvel indice de difficulté, basé cette fois sur le nombre de sujets ayant répondu à la question plutôt qu'au test, ne résout pas véritablement le problème. L'indice de difficulté risque d'être surestimé étant donné qu'il y a de fortes chances que ceux qui ont répondu à l'item soient les plus rapides et aussi les plus forts.

Lorsque l'indice de difficulté est calculé sur un item à choix de réponses, il faut tenir compte de la probabilité de réussir l'item sans vraiment connaître la réponse.

C'est ainsi qu'un item à réponse courte dont le coefficient de difficulté serait de 0,75 pourrait être considéré comme relativement facile. Ce ne serait pas le cas d'un item de type « vrai-faux » qui aurait un indice de difficulté de 0,75. Comme la probabilité de réussite au hasard est déjà de 0,50, l'item « vrai-faux » devrait être considéré comme relativement plus difficile que l'item à réponse courte.

Il est possible de corriger l'indice de difficulté pour l'effet du hasard chaque fois que l'on peut admettre que les leurres ont une chance à peu près égale d'être choisis. La formule de correction de l'indice de difficulté pour le hasard est la suivante :

$$p' = p - \left[\frac{1-p}{M-1}\right] \tag{6.2}$$

Dans l'équation (6.2), p' représente l'indice de difficulté corrigé, p représente l'indice de difficulté de départ et M le nombre de choix de réponses pour cet item.

Cette correction n'est pas nécessaire pour comparer les indices de difficulté d'un test constitué de questions semblables : par exemple, un ensemble de questions à quatre choix de réponse. On sait que, dans ce cas, la probabilité de réussite au hasard est de 1/M ou 0,25 pour toutes les questions. Par contre, si le format des questions varie (M = 2, 3, 4, 5), il sera nécessaire d'effectuer la correction pour pouvoir comparer la difficulté des items à partir d'une base commune.

Le tableau 2 illustre l'importance de cette correction lorsque l'on analyse des items comportant des nombres inégaux de choix de réponses. Dans ce tableau, l'item vrai-faux est réussi par une proportion plus grande d'élèves que les items à 3 ou à 5 choix de réponse. Toutefois, lorsque l'on applique la correction de l'équation (6.2), on se rend compte que ces trois items sont à toutes fins pratiques de mêmes degrés de difficulté. De plus, alors que la proportion de réussite p laisse entendre qu'il s'agit d'items réussis par la moitié au moins des élèves, la proportion p' révèle des items beaucoup plus difficiles, dont le pourcentage de difficulté se situe autour de 0,4.

Même si la correction pour l'effet du hasard comporte une certaine utilité dans l'analyse d'items, il n'est pas recommandé de l'appliquer systématiquement. Il y a plusieurs raisons pour ne pas effectuer une telle correction :

A. D'abord, il est très peu plausible qu'un sujet réponde véritablement au hasard. Celui-ci dispose toujours d'une connaissance partielle de la question qui lui permet d'éliminer des choix de réponses. Une question à cinq choix de réponses peut alors se ramener à trois ou à deux choix.

B. Ensuite, la correction pour l'effet du hasard ne change pas le classement des sujets. Le sujet le plus fort est toujours celui qui recevra le score le plus élevé et le plus faible est celui qui recevra le score le moins élevé. Seule la valeur absolue du score changera. Plutôt que de corriger le score individuel pour l'effet du hasard, il est préférable d'adapter le seuil de réussite en conséquence. Par exemple, un seuil de 0,60 pour un ensemble de questions « vrai ou faux » ne correspondrait pas à un seuil très élevé. Si l'on applique l'équation (6.2) à ce seuil, on découvre que le degré de difficulté corrigé pour l'effet du hasard est égal à 0,2. Un seuil de réussite de 20% n'est pas très exigeant.

Tableau 2 – Correction pour l'effet du hasard

Sujet#	Item Vrai/Faux	Item à 3 choix	Item à 5 choix
1	1	1	1
2	1	1	1
3	0	0	1
4	0	0	1
5	1	0	0
6	1	1	1
7	0	0	0
8	1	1	0
9	1	1	0
10	1	1	0
Difficulté P	0,70	0,60	0,50
Difficulté P'	0,40	0,40	0,38
Écart P-P'	0,30	0,20	0,13

1.2 Difficulté et distribution de l'item

Il existe un rapport étroit entre la difficulté d'un item et sa distribution. Lorsque l'item est soit trop facile, soit trop difficile, sa distribution devient asymétrique. Ce résultat est particulièrement évident dans le cas d'items dichotomiques. La figure 1 illustre ce rapport entre difficulté de l'item et symétrie de la distribution.

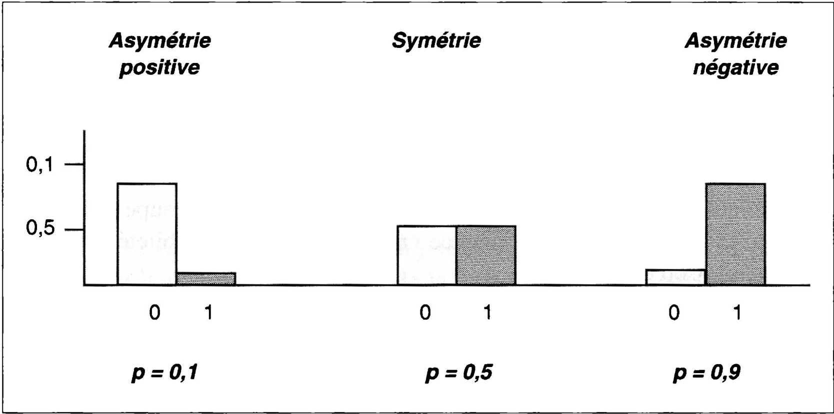


Figure 1 – Indice de difficulté et symétrie de la distribution des résultats à un item

Comme l'illustrent les trois distributions de la figure 1, les items trop faciles ou trop difficiles possèdent des distributions fortement asymétriques. Les items faciles permettent de bien discriminer parmi les sujets faibles et les items difficiles parmi les sujets forts. Si un sujet rate un item qui est réussi par 90% de ses pairs, cet échec est beaucoup plus grave que s'il avait raté un item réussi par 30% de ses pairs. C'est dans ce sens que l'on peut prétendre qu'un item facile permet de discriminer parmi les sujets faibles. Les sujets qui ratent ce genre d'item sont donc bien différents des autres. Le même raisonnement vaut pour les items difficiles. Les quelques sujets qui réussissent de tels items manifestent une habileté très supérieure à celle du groupe, pour autant qu'il ne s'agisse pas d'une réussite due au hasard. Les items difficiles permettent donc de sélectionner les meilleurs éléments d'un groupe. Par contre, les items de difficulté moyenne (p = 0,5) discriminent de manière symétrique : ils différencient aussi bien les sujets forts que les sujets faibles. C'est pourquoi cette catégorie d'items est particulièrement importante dans les évaluations où l'on souhaite différencier les sujets entre eux, peu importe le score total obtenu.

1.3 LA SÉLECTION DES ITEMS SELON LEUR DIFFICULTÉ

La difficulté des items a une influence importante sur le score total du test. C'est pourquoi le choix des items doit tenir compte de la proportion des répondants susceptibles de les réussir ou de les échouer. Que cette proportion soit estimée à partir d'un jugement d'expert avant l'examen ou qu'elle provienne des résultats d'un prétest, elle aura un impact sur notre capacité à discriminer au niveau du score total.

Prenons l'exemple de la figure 2. Les items y sont représentés par des cercles disposés sur une échelle de coefficients de difficulté de 1 à 0 (de facile à difficile). Chaque sujet y est représenté par une lettre associée à un score inscrit sur un drapeau placé à des points correspondants de l'échelle de difficulté des items. La figure 2 décrit la distribution des indices de difficulté des 16 items du test #1 et du test #2. Tous deux ont une caractéristique en commun : ils ne possèdent aucun item de difficulté intermédiaire. Comment des individus, assez forts pour réussir des items faciles mais trop faibles pour réussir des items très difficiles, seront-ils mesurés par ces tests ?

En fait, ni le test #1, ni le test #2 ne contribueront de façon significative à différencier de tels répondants, car aucun item ne fournit d'information à ce niveau. Dans le cas du test #1, l'éventualité la plus probable est qu'un sujet d'habileté intermédiaire (A) réussisse tous les items faciles et rate tous les items difficiles. Dans son cas, le score total ne dépend que des items faciles qu'il a réussis. Comme il n'y a aucun item de difficulté intermédiaire entre le groupe des items faciles et le groupe des items difficiles, le test #1 ne fera pas de différence entre deux sujets d'habileté intermédiaire, qu'ils soient en A ou en B.

Le test 2 ne fera pas davantage de distinction entre les deux sujets, mais il est susceptible de leur accorder un score total plus élevé car il comporte une plus grande proportion d'items faciles que d'items difficiles. Dans le test A, on retrouve 5/16 (31%) d'items faciles alors que dans le test B, cette proportion passe à 11/16 (69%).

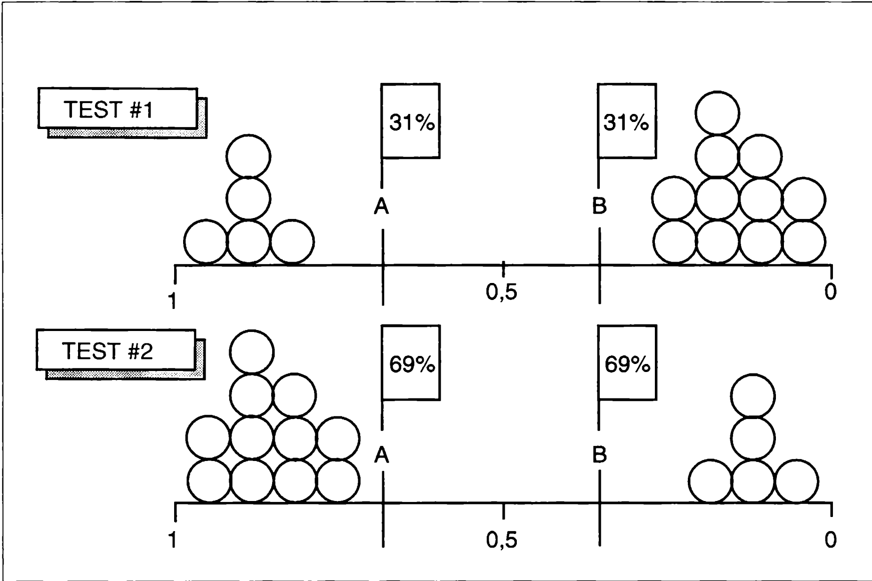


Figure 2 – Difficulté des items et discrimination entre les sujets

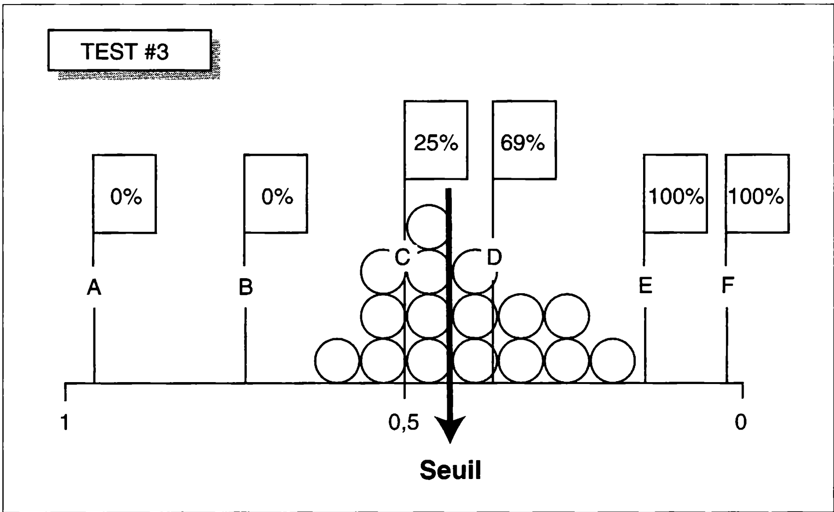


Figure 3 – Discrimination au seuil de réussite

La figure 3 représente une situation plus réaliste. On y constate un grand nombre d'items dont la difficulté est voisine de la note de passage. Cette note de passage représente le seuil au-dessus duquel on décidera, par exemple, de retenir une candidature pour un emploi, de classer un élève à un autre niveau ou de recommander une promotion. Une telle distribution des items permet d'accroître la différenciation entre les sujets qui obtiennent des résultats voisins de ce seuil d'exigence. À cause du grand nombre d'items dans le voisinage de la note de passage, de légères différences se tra-

duiront par des changements importants au niveau du score total. De cette façon, le score total du sujet nous permettra de bien discriminer entre ceux qui ont atteint et ceux qui n'ont pas atteint la valeur-seuil.

Le test #3 discrimine bien au seuil de réussite de 60%. Le sujet C en-dessous du seuil obtient un score bien différent (25%) du sujet D juste au-dessus (69%). Il y a en effet une grande proportion d'items qui mesurent l'habileté de ces deux sujets, ce qui se traduit par la possibilité de différences importantes au niveau du score total.

La situation n'est pas la même aux extrémités de la distribution pour les sujets A et B et pour les sujets E et F. A et B ne peuvent réussir que des items faciles, alors que E et F peuvent réussir virtuellement n'importe quel item, facile ou même très difficile. Comme il n'y a pas d'items ni très faciles, ni très difficiles, des sujets très faibles ou très forts sont susceptibles d'obtenir des résultats similaires. Mais dans une situation comme celle du test #3, est-il vraiment important de discriminer entre sujets qui auraient pu obtenir entre 90% et 100% ou entre 20% et 30% ? Sans doute non, puisque ces valeurs sont très différentes du seuil de passage et que dans chaque cas individuel, il est clair que les sujets ont satisfait ou non aux exigences minimales.

1.4 LA VARIANCE DE L'ITEM

Alors que le coefficient de difficulté nous indique dans quelle proportion un item est réussi, la variance de l'item nous renseigne à quel point les résultats à cet item sont dispersés ou non. Dans le cas d'items notés sur une échelle continue ou polychotomique, le calcul de la variance s'effectue au moyen de la formule habituelle (voir chapitre 1). Dans le cas d'items corrigés de façon dichotomique (0 ou 1), une formule simplifiée permet de calculer rapidement la variance. Elle est fournie par le produit de la proportion p des sujets ayant réussi l'item, par la proportion q des sujets l'ayant échoué (q valant $1\text{-}p$) :

$$\sigma_i^2 = p_i q_i \qquad (6.3)$$

Par exemple, pour calculer la variance de l'item dichotomique du tableau 1, on procède de la manière suivante :

$$s_i^2 = p_i q_i = 0,6 \times 0,4 = 0,24 \qquad (6.4)$$

La valeur maximale de variance pour un item corrigé de façon dichotomique égale 0,25. Cette valeur n'est possible que lorsque le coefficient de difficulté de l'item vaut 0,5. Donc, la dispersion d'un item dichotomique ne peut être maximale que lorsque la moitié des sujets ont réussi ou échoué l'item. Tout autre coefficient de difficulté donne lieu à moins de dispersion entre les sujets.

2. La discrimination de l'item

Lorsque l'on souhaite différencier entre eux les scores des sujets, la capacité de discrimination de l'item devient particulièrement importante. En effet, nous souhaitons retrouver dans un test des items qui contribuent à départager les sujets qui ont eu un

score total élevé à l'examen, des sujets qui ont eu un score total faible. Dans cette perspective, un « bon » item est un item qui serait réussi par une plus grande proportion de sujets ayant obtenu un score élevé à l'examen que par des sujets ayant obtenu un score faible. Une autre caractéristique de tels items est la suivante : il y a une forte corrélation entre la réussite à l'item et le score total au test.

Un test n'a pas toujours pour objectif de différencier les sujets entre eux. Au contraire, il existe de nombreuses situations d'évaluation où nous ne souhaitons pas qu'il y ait de différences entre les sujets. C'est le cas de la pédagogie de la maîtrise où un objectif doit être maîtrisé par une forte proportion des élèves (80% à 90%) avant de passer à l'objectif d'apprentissage suivant. Puisque l'intention est que tous les sujets atteignent l'objectif, la discrimination entre l'ensemble des élèves perd de son importance. Tout au plus, l'enseignant veut-il discriminer entre les sujets qui maîtrisent l'objectif et ceux qui ne le maîtrisent pas. Dans ce contexte, les items qui aideront le plus l'enseignant à faire cette distinction sont les items qui auront été le plus influencés par son enseignement. Ces items devraient être réussis par une forte proportion de sujets après l'enseignement et donner lieu à une distribution asymétrique négative des résultats. Les quelques sujets qui rateraient ce genre d'item sont ceux qui auraient besoin d'explications complémentaires ou d'un enseignement correctif. C'est le genre de discrimination que l'on veut obtenir en évaluation formative.

Il existe trois principaux types de discrimination que nous verrons dans les sections suivantes :

- l'indice de discrimination *D* ;
- les corrélations bisériales (r_{bis}) et de point -bisériales (r_{pbis}) ;
- l'indice de sensibilité à l'enseignement *S*.

2.1 L'INDICE DE DISCRIMINATION *D*

L'indice de discrimination *D* (Findley, 1956) est simplement la différence entre l'indice de difficulté d'un item pour le groupe dit « fort » (p_+) et l'indice de difficulté pour le groupe dit « faible » (p_-).

$$D = p_+ - p_- \quad (6.5)$$

Plus l'écart *D* est grand, plus l'item discrimine entre les sujets ayant eu un score total élevé et les élèves ayant eu un score total faible.

Suite à la proposition de Kelley (1939), le groupe fort est constitué de ceux qui ont obtenu un score total qui les situe dans la catégorie des 27% supérieurs et le groupe faible dans la catégorie des 27% inférieurs. Par exemple, dans un groupe comptant 30 répondants, on prendra les huit ($0,27 \times 30 = 8,1$) résultats les plus élevés et les huit résultats les plus bas pour calculer les deux indices de difficulté p_+ et p_-

L'indice *D* peut prendre n'importe quelle valeur entre -1 et +1. Une valeur 0 signifie qu'un item est tout aussi bien réussi par les sujets qui ont eu un score total élevé que par les sujets qui ont eu un score total faible. Une valeur négative signifie que l'item a été réussi par une plus grande proportion de sujets qui ont eu un score total peu élevé à l'ensemble du test. De telles valeurs soulèvent des doutes quant à l'opportunité

de conserver ce genre d'item dans le calcul du résultat total. Ebel (1965) propose les valeurs repères suivantes pour interpréter le coefficient de discrimination *D* :

- 0,40 et plus — item qui discrimine très bien ;
- 0,30 à 0,39 — item qui discrimine bien ;
- 0,20 à 0,29 — item qui discrimine peu ;
- 0,10 à 0,19 — item-limite, à améliorer ;
- Moins de 0,10 — item sans utilité réelle pour l'examen.

L'indice *D* est particulièrement utile pour le calcul manuel de la discrimination. En effet, il ne porte que sur la moitié (54%) des données, ce qui diminue le travail de calcul. De plus, il donne des résultats fort semblables à ceux des méthodes corrélationnelles plus complexes. L'indice *D* convient donc tout à fait à l'analyse d'items de tests scolaires, à condition que les items soient suffisamment nombreux (30 ou plus). Lorsque le nombre d'items est restreint, l'indice de discrimination est artificiellement gonflé du fait que chaque item compte pour une proportion trop importante du score total.

Le tableau 3 présente un cas pratique de calcul de *D* par la méthode de pointage. Dans une classe de 33 élèves, il y aura 9 élèves dans le groupe fort et 9 élèves dans le groupe faible. Un simple pointage des questions réussies par les élèves de chacun de ces groupes — sur un copie vierge de l'examen ou sur le solutionnaire — permet de repérer en un coup d'oeil les items qui discriminent bien de ceux qui ne discriminent pas. Par exemple, les items 1 et 5 discriminent très bien. Les items 2 et 4 discriminent faiblement car ils sont presqu'aussi bien réussis dans chaque groupe. Enfin, l'item 3 présente un problème sérieux : il s'agit d'un item très difficile réussi par un seul élève appartenant au groupe faible. Il pourrait s'agir d'une réussite due à la chance, surtout s'il s'agit d'une question à choix de réponses.

Tableau 3 – Calculs des coefficients de difficulté et de discrimination (n=33)

Questions #	Groupe fort (/9)	Groupe faible (/9)	P	D
1. ____________	✔✔✔✔✔	✔✔✔	10/18	4/9
____________ ?	✔✔		*(0,56)*	*(0,44)*
2. ____________	✔✔✔✔✔	✔✔✔✔✔	15/18	1/9
____________ ?	✔✔✔	✔✔	*(0,83)*	*(0,11)*
3. ____________		✔	1/18	-1/9
____________ ?			*(0,06)*	*(-0,11)*
4. ____________	✔✔✔	✔	4/18	2/9
____________ ?			*(0,22)*	*(0,22)*
5. ____________	✔✔✔✔✔	✔	9/18	7/9
____________ ?	✔✔✔		*(0,50)*	*(0,78)*

En plus du pointage, le tableau 3 présente les résultats du calcul des indices *p* et *D*. Dans le cas particulier de l'indice *p*, une méthode approximative a été employée qui diminue la quantité de calculs. Alors que *D* est calculé par la différence de difficulté de chaque item pour chaque groupe, *p* est calculé en faisant la moyenne de ces difficultés. Cette valeur est généralement une très bonne approximation de la valeur de *p* calculée pour l'ensemble des sujets. Il est donc relativement simple, avec 18 élèves sur 33, de calculer l'indice de difficulté et l'indice de discrimination pour tous les items en une seule opération rapide.

Lorsque des items discriminent peu ou encore discriminent négativement, il peut être nécessaire d'étudier ces items de plus près afin de mieux comprendre ce qui a pu se passer. Dans le cas de questions à choix de réponses, il est possible de considérer quel pourcentage d'élèves du groupe fort et du groupe faible a opté pour chacun des choix de réponse. À partir de ces résultats, on peut alors calculer un indice de discrimination non seulement pour la bonne réponse, mais aussi pour les leurres. Ces coefficients de discrimination pour les leurres devraient être tous négatifs, car ils sont censés être choisis par une plus grande proportion d'élèves du groupe faible que du groupe fort.

Le tableau 4 décrit l'analyse d'un item à quatre choix de réponses. L'item ne discrimine pas. En effet, l'indice *D* est nul car la bonne réponse *(b)* a été choisie par autant d'élèves du groupe fort que du groupe faible. C'est plutôt le leurre *(c)* qui permet de discriminer entre ces deux groupes. L'indice *D* pour ce leurre est positif et relativement élevé (0,33). En fait, plus d'élèves du groupe fort ont choisi cette option de préférence à la bonne réponse. Les deux autres leurres semblent fonctionner de manière plus ou moins adéquate : *(d)* est un leurre attirant choisi par deux fois plus d'élèves du groupe faible et *(a)* n'est pas un leurre très attirant puisqu'aucun élève du groupe faible ne l'a choisi.

Tableau 4 – Discrimination des choix de réponses

Questions #	Groupe fort (/9)	Groupe faible (/9)	P	D
1. ______________ ______________?				
a) ______________		✔	1/18 *(0,06)*	-1/9 *(-0,11)*
b) ______________	✔✔✔	✔✔✔	6/18 *(0,33)*	0/9 *(0,00)*
c) ______________	✔✔✔✔	✔	5/18 *(0,28)*	3/9 *(0,33)*
d) ______________	✔✔	✔✔✔✔	6/18 *(0,33)*	-2/9 *(-0,22)*

Face à des résultats tels que ceux du tableau 4, on peut se demander si l'option *(c)* n'était pas une réponse acceptable ou s'il n'y a pas eu d'erreur dans la clé de correction. Si ces explications ne conviennent pas, il serait utile de découvrir les raisons pour lesquelles le choix *(c)* a été si attirant lors de la discussion des résultats avec les élèves, dans le cas d'un examen de rendement scolaire.

2.2 Les indices corrélationnels de discrimination

La micro-informatique étant de plus en plus répandue, les constructeurs de test ont maintenant à leur disposition des logiciels qui effectuent l'analyse des résultats. Plusieurs de ces logiciels fournissent une analyse d'items comprenant le calcul d'un indice de discrimination. Cet indice de discrimination porte sur l'ensemble des données et repose sur le calcul d'une corrélation entre le score à l'item et le score total à l'examen. La corrélation de Pearson, décrite au chapitre 2, permet de calculer de tels indices.

Le calcul de la corrélation de Pearson requiert des échelles continues de mesure. Lorsque l'item est corrigé de manière dichotomique (0, 1) ou encore de manière ordinale (A, B, C, D, E ou encore 0, 1, 2, 3 et 4 points), le *r* de Pearson ne fournit pas une valeur exacte de la corrélation entre deux variables.

Encadré 1

Méthodes alternatives de calcul de la corrélation

Échelles de mesure	Dichotomique	Dichotomisée	Continue
Dichotomique	ϕ	ϕ_{bis}	r_{pbis}
Dichotomisée		r_{ter}	r_{bis}
Continue			r r_s

Il existe plusieurs alternatives au *r* de Pearson. Elles sont décrites dans l'encadré 1. Le choix entre chacune de ces méthodes dépend des postulats que l'on fait sur la nature de l'échelle de mesure employée pour chacune des deux variables en corrélation. Trois catégories d'échelle sont prises en ligne de compte : l'échelle dichotomique, l'échelle dichotomisée et l'échelle continue.

Ces considérations sur la nature de l'échelle de scores sont importantes pour choisir la méthode corrélationnelle la plus appropriée au calcul de la discrimination

ainsi que d'autres indices nous permettant d'approfondir notre analyse des items à l'examen. Il est possible de résumer ces considérations aux cinq points suivants :

1. Lorsque les deux variables sont continues, le *r* de Pearson doit être utilisé ; lorsque l'une des deux variables est ordinale et ne se distribue pas normalement, il est préférable d'utiliser le r_S de Spearman.

2. Lorsque l'une des variables est continue et que l'autre variable est réellement dichotomique (telle que le sexe), le calcul de la corrélation de Pearson peut s'effectuer au moyen du coefficient de corrélation point-bisériale. Cependant, la valeur maximale de 1 ne peut être atteinte que lorsque la variable dichotomique est symétrique, c'est-à-dire qu'il y a un nombre égal de sujets dans chaque catégorie dichotomique. Dans un cas extrême où 95% des sujets tombent dans l'une des deux catégories, Lord et Novick (1968) ont démontré que la valeur du r_{pbis} variait entre -0,5 et +0,5.

3. Lorsque l'une des variables est continue et que l'autre variable est une variable continue dichotomisée (telle qu'un item corrigé 0,1), la corrélation bisériale fournit une estimation de la corrélation de Pearson qui aurait pu être obtenue si la seconde variable n'avait pas été dichotomisée.

4. Lorsque les deux variables sont réellement dichotomiques, le calcul de la corrélation de Pearson peut être remplacé par celui de la corrélation ϕ *(phi)*. Cependant, comme dans le cas de la corrélation point-bisériale, la valeur maximale de 1 ne peut être atteinte que lorsque les deux variables sont symétriques, c'est-à-dire que la moitié des sujets se retrouve dans chaque catégorie.

5. Lorsque les deux variables sont dichotomisées, le calcul de la corrélation tétrachorique est préférable. Le calcul de cette corrélation est complexe et difficilement réalisable, même avec les logiciels disponibles (voir Dixon, 1981). Dans le cas des items de difficulté moyenne, la corrélation ϕ et la corrélation tétrachorique fournissent les mêmes résultats. La différence est plus importante dans les cas extrêmes où des items très faciles ou très difficiles sont mis en corrélation. Le calcul des corrélations tétrachoriques est particulièrement recommandé si l'on souhaite réaliser une *analyse factorielle* sur la matrice des intercorrélations entre les items. Mis à part ce cas bien particulier, il semble qu'à défaut de pouvoir employer les corrélations tétrachoriques, les corrélations ϕ peuvent constituer une alternative pratique, quoiqu'imparfaite.

L'encadré 2 fournit un exemple de calcul du coefficient ϕ. Par exemple, il pourrait s'agir de déterminer la corrélation entre le fait d'avoir choisi « vrai ou faux » à une question et « vrai ou faux » à une deuxième question. Dans l'exemple, p_j et $_{pk}$ représentent les proportions de ceux qui ont répondu « vrai » aux items *j* et *k* respectivement, alors que q_j et q_k représentent les proportions de ceux qui ont répondu « faux » à ces deux items. Enfin, p_c représente la proportion de ceux qui ont répondu « vrai » aux deux items. La corrélation trouvée entre les deux items est relativement faible, considérant le petit nombre de sujets sur lequel se fonde la corrélation (voir section 3.2 de ce chapitre pour un test de signification sur les valeurs de corrélation).

Encadré 2

Corrélation ϕ

Données

j	k
1	1
1	0
0	0
1	0
0	0

$P_{jk} = 0,2$

$P_j = 0,6$ $P_k = 0,2$

$Q_j = 0,4$ $Q_j = 0,8$

Équation

$$\phi_{jk} = \frac{p_{jk} - p_j\ p_k}{\sqrt{p_j q_j - p_k q_k}}$$

Calcul

$$\phi_{jk} = \frac{0,2 - 0,6 \times 0,2}{\sqrt{0,6 \times 0,4 \times 0,2 \times 0,8}} = 0,41$$

L'encadré 3 illustre comment se calcule la corrélation point-bisériale. Supposons qu'il s'agisse ici de déterminer si le fait d'être un homme ou une femme permet de différencier les sujets quant au score total *X*. La variable dichotomique *i* est ici le sexe, où *p* représente la proportion de femmes et *q* la proportion de hommes. La variable continue est *X* où représentent respectivement la moyenne et l'écart type des résultats de tous les élèves, hommes et femmes, au score total *X* et où représente la moyenne des femmes seulement (*k=1* dans ce cas-ci). Finalement, *p* représente la proportion de femmes *(k=1)* et *q* la proportion de hommes *(k=0)*. Notez bien : *p* représente toujours la proportion des sujets dont les scores entrent dans le calcul de .

Encadré 3

Corrélation point-bisériale

Données

i	X
1	5
1	2
0	1
1	5
0	2

$p = 0,6$ $\bar{X}_+ = 4$

$q = 0,4$ $\bar{X} = 3$

(Les valeurs encadrées 5, 2 et 5 sont reliées à +.)

Équation

$$r_{pbis} = \frac{(\bar{X}_+ - \bar{X})}{s_X} \sqrt{\frac{p}{q}}$$

Calcul

$$r_{pbis} = \frac{(4 - 3)}{1,67} \times 1,22 = 0,73$$

Dans ce cas-ci, une corrélation modérée (0,73) indique que les femmes réussissent mieux au test que les hommes et qu'il y a une association entre le sexe du sujet et sa probabilité de réussir le test. Si cette corrélation indiquait une différence réelle entre hommes et femmes, elle signifierait que le test mesure une habileté où les femmes sont généralement meilleures. Toutefois, il se pourrait aussi qu'une telle corrélation soit le fruit d'une mauvaise sélection des items : elle indiquerait alors un biais dans le choix des items qui défavoriserait systématiquement les hommes. Ce genre de préoccupation est particulièrement important dans les tests nationaux et internationaux dont les résultats peuvent servir à prendre des décisions importantes sur les programmes d'étude ou sur l'avenir des candidats (voir section 6 sur l'étude du biais des items).

L'encadré 4 présente un exemple de calcul de corrélation bisériale sur des données identiques à celles de l'encadré 3. Dans ce cas-ci, nous considérons que la variable *i* n'est pas une variable dichotomique comme le sexe, mais une variable dichotomisée, tel que la réussite ou l'échec à un item. Plutôt que de chercher à déterminer si le score total permet de discriminer entre hommes et femmes, comme dans le cas précédent, nous chercherons à savoir si l'item permet de discriminer entre ceux qui ont eu un score total élevé et ceux qui ont eu un score faible au test.

Encadré 4

Corrélation bisériale

Données

i	X
1	5
1	2
0	1
1	5
0	2
p = 0,6	$\bar{X}_+ = 4$
q = 0,4	$\bar{X} = 3$

Équation

$$r_{pbis} = \frac{(\bar{X}_+ - \bar{X})}{s_X} \sqrt{\frac{p}{Y}}$$

Calcul : Y = 0,387

$$r_{pbis} = \frac{(4 - 3)}{1,67} \times 1,55 = 0,93$$

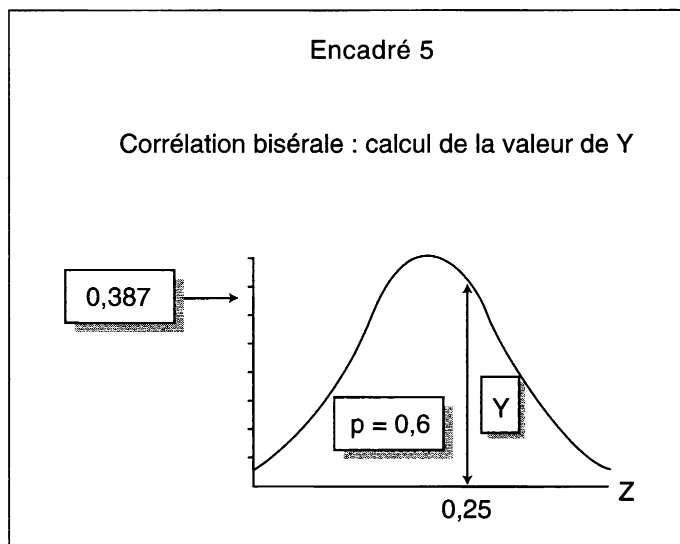
Dans cet exemple de calcul, la signification des symboles est la même que dans le cas de la corrélation bisériale, à une importante exception près : le calcul de la valeur de Y. Celle-ci correspond, comme l'indique l'encadré 5, à la hauteur de la courbe normale au point *z* correspondant à une densité de probabilité égale à *p*. Dans notre exemple, la valeur de *p* est de 0,6. Dans une distribution normale centrée réduite, une telle densité de probabilité correspond à un score z = 0,25. En effet, selon la distribution des probabilités de la loi normale, il y a six chances sur 10 d'obtenir un score égal ou supérieur à 0,25 écart type au-dessus de la moyenne. Cette probabilité nous est fournie par une table des valeurs de la loi normale. Cette même table nous fournit également la valeur de la hauteur de la courbe normale au point z = 0,25. Pour cette valeur de *z*, *Y = 0,387.*

Dans notre exemple, une corrélation de 0,93 signifie que l'item *i* permet de bien discriminer les sujets forts des sujets faibles. Il s'agit là d'un item à conserver si notre intention est de discriminer entre les personnes.

Lord et Novick (1968) ont démontré que la corrélation bisériale obtenue est 20% supérieure au coefficient de corrélation point-bisériale. L'équation (6.6) permet de transformer une corrélation point-bisériale en corrélation bisériale.

$$r_{bis} = \frac{\sqrt{pq}}{Y} r_{pbis} \qquad (6.6)$$

Dans le cas de valeurs extrêmes de p ou q, Magnusson (1967) a démontré que la corrélation bisériale pouvait être jusqu'à quatre fois supérieure à la corrélation point-bisériale. Ceci est dû au fait que la faible variance des items affecte grandement la valeur maximum que peut prendre la corrélation point-bisériale, qui est un équivalent algébrique du r de Pearson. Il est donc primordial, lorsque l'on utilise un logiciel quelconque de calcul des indices de discrimination, de savoir quel type de corrélation est employé pour calculer la corrélation item-total. Enfin, en comparant les résultats publiés sur les analyses d'items de tests commerciaux, il faut également tenir compte que des indices de discrimination rapportés en corrélations bisériales seront toujours plus élevés que les corrélations point-bisériales, particulièrement dans le cas de valeurs extrêmes de p ou q.



L'encadré 6 fournit un exemple de calcul de la corrélation par rangs de Spearman. Kendall (1938) a également proposé une formule de calcul de la corrélation par rangs, mais celle-ci fournit des résultats numériquement très différents de ceux de Pearson, ce qui les rend difficilement comparables. La formule de Spearman requiert que les résultats soient d'abord transformés en rangs et qu'un écart entre les rangs occupés par la même personne sur les deux variables soit calculé. Si une personne est la première sur l'une des deux variables, elle devrait être la première sur l'autre variable si celles-ci sont effectivement en corrélation.

Dans l'exemple de l'encadré 6, la variable *X* représente le résultat d'un élève à une question à réponse élaborée corrigée sur 10 points et la variable *Y* représente le résultat à une question corrigée sur 20 points. Une forte corrélation entre ces deux questions indiquerait que le correcteur a fait preuve d'une certaine cohérence dans la correction et/ou que les deux questions mesurent une caractéristique relativement homogène.

Selon Hotelling et Pabst (1936), la corrélation de Spearman possède une efficacité relative de 91% par rapport à la corrélation *r* de Pearson. Ceci signifie qu'une corrélation par rangs estime la corrélation entre deux variables mesurées sur un échantillon de 100 sujets avec la même précision qu'une corrélation de Pearson portant sur 91 sujets lorsque les conditions pour le calcul d'une corrélation de Pearson sont respectées. L'avantage particulier de la corrélation de Spearman est de permettre une bonne estimation de la corrélation lorsque les postulats de base de la corrélation de Pearson ne tiennent pas, comme c'est le cas lors d'une distribution de rangs. Elle est donc recommandée chaque fois que l'une des deux variables ne se distribue pas normalement ou encore ne rencontre pas les conditions d'une échelle à intervalles égaux.

Les coefficients de corrélation par rangs sont particulièrement utiles lorsque l'on veut s'assurer du degré de concordance entre juges. Deux juges qui n'ordonneraient pas les sujets de la même manière lors d'une compétition, ne contribueraient pas à départager un vainqueur. La corrélation de Spearman est tout à fait indiquée lorsque l'on cherche à déterminer le degré de concordance entre juges pris deux à deux. Lorsque l'on veut évaluer le degré global de concordance entre plus de deux juges, le *W de Kendall* (Siegel & Castellan, 1988) — une autre mesure de corrélation par rangs — permet d'estimer au moyen d'une seule valeur, à quel point chaque juge diffère du rang moyen octroyé par l'ensemble des juges (voir chapitre 5 sur la validité)

Encadré 6

Corrélation par rangs de Spearman

Données

X → rang	Y → rang	D(i)
5 → 4	15 → 4	0
1 → 1	6 → 1	0
3 → 3	12 → 2	1
2 → 2	13 → 3	-1
7 → 5	19 → 5	0

$\sum D_i^2 = 2$

Équation

$$r_s = 1 - \frac{6 \sum D_i^2}{n^3 - n}$$

Calcul

$$r_s = 1 - \frac{6 \times 2}{125 - 5} = 0,90$$

2.3 Le choix du bon indicateur de discrimination

Il existe donc une grande variété d'indicateurs corrélationnels s'ajoutant à l'indice de discrimination *D* pour déterminer si un item permet de différencier les sujets obtenant un score total élevé de ceux obtenant un score faible. Plusieurs recherches ont démontré une forte corrélation entre ces indicateurs (Englehart, 1965 ; Beuchert et Mendoza, 1979 ; Findley, 1956 ; Oosterhof, 1976). Les plus importantes différences se produisent pour les items dont les coefficients de difficulté comportent une valeur extrême (près de 0 ou de 1).

Crocker et Algina (1985, p. 319) ont formulé cinq recommandations pour faciliter le choix des indices de discrimination disponibles pour items dichotomiques :

1. Lorsque les items sont de difficulté modérée, l'ensemble des méthodes se valent. Les méthodes corrélationnelles ont cependant l'avantage de permettre un test de signification statistique. Un tel test n'existe pas pour l'indice *D*.
2. Lorsque l'objectif est de choisir parmi des items se situant à chaque extrémité du spectre des niveaux de difficulté, la corrélation bisériale devrait être employée.
3. Si l'on suspecte que les futurs échantillons de sujets auxquels sera administré le test seront d'habiletés fort différentes des sujets sur lesquels le test a été mis à l'essai, il est préférable d'utiliser la corrélation bisériale.
4. Si l'on s'attend à ce que le test soit utilisé avec des sujets de même niveau d'habileté, la corrélation point-bisériale semble la mieux indiquée.
5. Lorsque l'item et la variable critère sont cotés de manière dichotomique (c'est le cas lorsque le score total est transformé en « maîtrise »/« non maîtrise »), le coefficient tétrachorique devrait être employé surtout si item et critère prennent des valeurs extrêmes. Toutefois, il est très difficile de calculer cette valeur et plusieurs s'accommoderont du coefficient ϕ.

3. Rapport entre difficulté et discrimination de l'item

Peu importe le type d'indicateur de discrimination employé, lorsque l'item est trop facile ou trop difficile, l'estimation de sa contribution à la différenciation des sujets au niveau du score total devient risquée. Tant l'indice *D* que les indices corrélationnels sont, en effet, influencés par la difficulté de l'item.

Parfois, les constructeurs de tests sont placés face à un dilemme. D'une part, ils veulent obtenir un score total qui leur permette de différencier les sujets. D'autre part, ils ne veulent pas renoncer à poser des questions faciles ou difficiles, car ils permettent de discriminer les sujets qui se situent aux extrémités de la distribution de l'aptitude mesurée. Nous avons vu dans la section 1.1 que les items difficiles, même s'ils ne permettent pas de discriminer adéquatement parmi tous les sujets, favorisent une meilleure discrimination parmi les sujets forts. De même, les items faciles nous permettent de bien discriminer parmi les sujets faibles.

Qu'en est-il lorsque nous souhaitons discriminer aussi bien parmi les sujets forts que parmi les sujets faibles, comme c'est souvent le cas en éducation lors de

l'évaluation sommative ou en psychométrie avec les tests d'aptitude ? Dans de tels cas, les items faciles ou difficiles jouent un rôle plus complexe et c'est au concepteur de s'interroger sur ce rôle en fonction des objectifs d'évaluation. L'analyse d'items peut l'aider à se poser les questions pertinentes quant au rôle joué par chaque item ainsi que sur les moyens appropriés pour améliorer la qualité de l'instrument de mesure.

3.1 Le choix du « bon » item

Un « bon » item nous permet d'atteindre notre objectif d'évaluation. Cet objectif sera atteint en choisissant des items de difficulté et de discrimination adéquates. La figure 4 illustre comment peut s'effectuer la sélection des meilleurs items en fonction de leur difficulté et de leur discrimination (indice *D*).

L'abscisse de la figure 4 représente le coefficient de difficulté de l'item : elle ne peut prendre que des valeurs positives de 0 à 1. L'ordonnée permet de situer les items en fonction de leur indice de discrimination *D* : elle peut prendre des valeurs de -1 à +1. Étant donné la relation entre *D* et *p*, certaines combinaisons de valeur sont impossibles : un item trop facile ou trop difficile ne peut avoir une valeur de discrimination élevée. C'est ce qu'indiquent les barres obliques en tirets. Elles délimitent les régions du plan cartésien où les couples de valeurs *p* et *D* ne peuvent se produire. Enfin, une série de droites horizontales indiquent les seuils critiques de *D* suggérés par Ebel (1965).

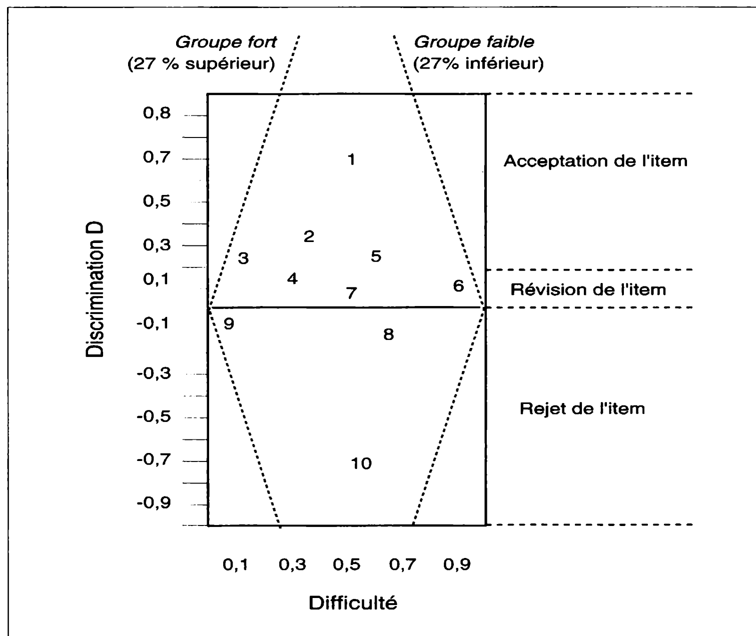


Figure 4 – Rapports entre discrimination et difficulté de l'item

Une fois que nous prenons en considération l'ensemble de ces facteurs, il est possible de mieux interpréter la signification des valeurs de *p* et de *D* pour chaque

item. La figure 4 présente 10 combinaisons différentes de difficulté et de discrimination d'items. Voici quelques interprétations possibles de chacune de ces 10 situations :

1. L'item 1 devrait être retenu. Il représente l'item idéal pour différencier les sujets : difficulté moyenne et discrimination élevée.
2. L'item 2 mérite aussi d'être retenu. C'est un item légèrement difficile, mais qui discrimine assez bien. Il se situe au-dessus du seuil recommandé par Ebel où une révision serait nécessaire.
3. L'item 3 se situe dans la zone de révision. Il discrimine peu, mais il faut tenir compte que c'est également un item très difficile. En fait, sa valeur de discrimination se situe très près du maximum possible à ce niveau de difficulté. Faut-il alors vraiment réviser cet item ? Non, car cet item nous permet de discriminer au maximum de ce à quoi l'on peut s'attendre à ce niveau.
4. L'item 4 mérite une attention particulière. Il se situe dans la zone de révision et de plus il est très en deçà du maximum qu'il peut atteindre.
5. L'item 5 présente un cas similaire à l'item 4. De degré de difficulté moyenne, il n'a qu'une faible discrimination. S'il s'agit d'un item à choix de réponses, il serait intéressant de revoir la distribution des choix de réponses de chaque leurre, ainsi que de calculer un coefficient de discrimination par leurre (voir section 2.1).
6. Si l'on ne se fiait qu'à la discrimination, l'item 6 devrait être rejeté immédiatement. C'est un item qui ne peut discriminer car il est réussi par la presque totalité des répondants (90% et plus). Ce n'est pas l'item qu'il faut revoir, mais plutôt l'opportunité de l'inclure. Si l'on souhaite mesurer des prérequis jugés essentiels ou l'atteinte d'objectifs minima, alors cet item mérite d'être conservé. Il nous faut accepter cependant qu'un tel item ne nous permettra pas de différencier tous les sujets, mais qu'il pourra nous être fort utile pour identifier les sujets les plus faibles.
7. L'item 7 ne sert à rien. Il ne discrimine pas du tout parmi les sujets malgré qu'il s'agisse d'un item de difficulté moyenne. On ne peut donc imputer sa faible discrimination au fait qu'il soit trop facile ou trop difficile. Il devrait être éliminé car, avec ou sans cet item, les résultats des sujets ne sont guère différents.
8. L'item 8 mérite également d'être retiré du test. C'est un cas similaire à l'item 7 avec un inconvénient en plus : s'il est conservé, il diminuera les différences entre les sujets, car il discrimine de manière négative. Cet item envoie donc un message contradictoire par rapport au message envoyé par l'ensemble des items du test.
9. L'item 9 est un cas particulier de discrimination négative. C'est un item très difficile qui est mieux réussi par les sujets qui ont les moins bons résultats au test. Il peut s'agir de quelques sujets qui ont répondu au hasard.
10. L'item 10 est un cas grave de discrimination négative. De difficulté moyenne, il est, comme l'item 9, réussi par un plus grand nombre de sujets du groupe « faible ». À la différence de l'item 9, il n'est pas possible d'attribuer un tel résultat à une réussite au hasard, car il ne s'agit pas ici de quelques patrons de

réponse aberrants. Ce genre d'item suggère plutôt une erreur dans la grille de correction ou encore un problème dans l'apprentissage antérieur des sujets.

La figure 4 nous permet d'articuler indices de difficulté et de discrimination dans l'analyse des items à un test. Quoique l'exemple fourni vaille pour l'indice de discrimination *D*, le même type d'analyse peut être réalisé avec les indices corrélationnels. Dans ce cas, les valeurs maximales de corrélation changent également en fonction de l'indice de discrimination et un test de signification sur les valeurs de corrélation remplace les seuils pratiques déterminés par Ebel (1965).

3.2 Test de signification des indices corrélationnels de discrimination

Lorsqu'un indice corrélationnel de discrimination est faible, il est important de déterminer si la corrélation entre l'item et le score total est significativement différente de 0 ou si elle aurait pu être obtenue au hasard. Lorsque le nombre de sujets est supérieur à 50, Magnusson (1967) a démontré que l'écart type de la distribution des *r* de Pearson autour d'une moyenne de 0 était estimé par l'équation suivante :

$$s_r = \frac{1}{\sqrt{N-1}} \qquad (6.7)$$

où s_r est l'écart type de la distribution de *r* et *N* le nombre de sujets ayant servi au calcul de la corrélation.

De l'équation (6.7) on peut retenir que plus l'échantillon est petit, plus grande devra être la corrélation entre deux variables avant que celle-ci ne puisse être considérée comme significativement différente de 0. Plus le nombre de répondants à un test est petit, plus l'indice de discrimination devra être grand avant que l'on considère qu'un item contribue à différencier les sujets quant à leur score total.

Le même écart type est utilisé pour déterminer le degré de signification des corrélations point-bisériale et ϕ. Dans le cas de la corrélation bisériale, l'écart type de la distribution est fourni par la formule développée par Kurtz et Mayo (1979) :

$$s_{r_{bis}} = \frac{\sqrt{pq/(N-1)}}{Y} \qquad (6.8)$$

où *Y* représente la valeur de l'ordonnée de la courbe normale au point *z* correspondant à une densité de probabilité *p* (voir encadré 5) ; *p* est la proportion d'élèves qui a réussi l'item ; *q* représente la proportion d'élèves qui a échoué l'item *(1-p)*. Enfin, *N* représente le nombre d'élèves ou de couples d'observations.

Dans tous les cas, l'écart type calculé par les équations (6.7) ou (6.8) sert à déterminer un intervalle de confiance de 95% ou 99% autour de la moyenne 0. Si la valeur de corrélation calculée se situe à l'intérieur de cet intervalle, c'est qu'elle n'est pas significativement différente de 0 au seuil de signification choisi (0,05 ou 0,01).

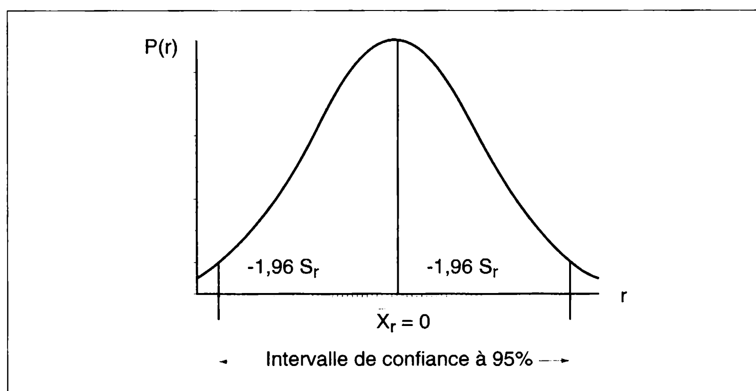


Figure 5 – Intervalle de confiance à 95% d'une valeur de corrélation

La figure 5 fournit un exemple d'un test de signification d'une valeur de corrélation. Supposons que nous soyons intéressés à déterminer à partir de quelles valeurs une corrélation calculée sur un échantillon de 82 sujets est significativement différente de 0. Nous devons d'abord estimer la valeur de dispersion des corrélations autour de *r = 0* selon l'équation (6.9) :

$$s_r = \frac{1}{\sqrt{82-1}} = \frac{1}{9} = 0{,}11 \qquad (6.9)$$

Les valeurs comprises entre $\pm\, 1{,}96 s_r$ déterminent un intervalle de confiance à l'intérieur duquel se situent 95% des valeurs de corrélations qui peuvent se produire au hasard entre 82 couples de données pour lesquels il n'y a pas de corrélation. Si la valeur observée de corrélation excède les limites de cet intervalle, alors nous pouvons la considérer comme significativement différente de 0 avec un risque d'erreur de type I de 0,05. Dans le cas qui nous intéresse, cet intervalle est compris entre ± 0,22. Une corrélation de 0,34 serait donc considérée comme significativement différente de 0.

3.3 Calculs pratiques des indices de difficulté et de discrimination

Lors de l'analyse des items d'examens de rendement scolaire, il n'est pas toujours nécessaire d'employer tout l'arsenal des outils statistiques à notre disposition. De plus, il n'est pas toujours possible, ni simple, d'avoir recours à un programme d'ordinateur. Enfin, l'analyse des items nécessite que les données de chaque individu soient entrées pour chaque item, ce qui peut représenter une tâche considérable.

Pour l'avenir prévisible, il y a de bonnes raisons de croire que l'analyse des résultats à un examen se fera encore de façon artisanale. Toutefois, elle peut être plus efficace si nous savons exploiter les rapports qui existent entre les principaux indicateurs statistiques. C'est ainsi que nous avons démontré dans la section 2.1 que nous pouvions estimer l'indice de difficulté et l'indice de discrimination à partir des résul-

tats d'environ la moitié des sujets. Ceci constitue un allégement important des efforts de calcul.

Il est possible d'aller encore plus loin et d'utiliser les propriétés des indices de discrimination pour estimer l'écart type des résultats et la cohérence interne d'un test. En effet, plus la discrimination moyenne des items est élevée, plus on peut s'attendre à une grande dispersion des résultats et à une forte intercorrélation entre les items, comme nous l'avons démontré au chapitre 4, section 1.

Wiersma et Jurs (1990) ont démontré que la somme des indices de discrimination est environ 2,45 fois plus grande que l'écart type des scores totaux. On peut donc estimer l'écart type au moyen de l'équation suivante :

$$s_X \cong \frac{\sum D}{2{,}45} \tag{6.10}$$

De la même manière, Wiersma et Jurs proposent les alternatives suivantes au calcul des valeurs de *KR20* et *KR21* :

$$KR20 = \frac{j}{j-1}\left[1 - \frac{6\sum pq}{y(\sum D)^2}\right] \tag{6.11}$$

$$KR21 = \frac{j}{j-1}\left[1 - \frac{6\bar{X}(j-\bar{X})}{j(\sum D)^2}\right] \tag{6.12}$$

Dans les trois équations précédentes, j représente le nombre d'items. Pour calculer l'écart type, nous n'avons besoin que de la somme des indices de discrimination D. Pour calculer *KR20* et *KR21*, nous avons besoin aussi de la moyenne des scores totaux ($\bar{X}$). En utilisant les formules précédentes, nous pouvons donc réaliser une analyse fort complète des résultats à un examen à partir des résultats de la moitié des élèves seulement.

4. Indices de discrimination pour la mesure critériée

Les indices de discrimination que l'on vient de voir conviennent particulièrement aux tests qui ont pour objectif de différencier les répondants entre eux. Ce n'est pas l'objectif poursuivi par tous les tests. Dans le cadre d'une pédagogie de maîtrise ou d'une évaluation formative, nous ne nous attendons pas à ce que notre instrument de mesure discrimine également entre tous les sujets. Par contre, nous voulons savoir s'il permet de faire la différence entre les élèves qui maîtrisent ou qui ne maîtrisent pas un objectif au seuil de réussite fixé à l'avance.

Les items les plus utiles en mesure critériée sont ceux qui sont les plus sensibles à l'enseignement. Si l'enseignement a été profitable, le degré de difficulté de ces items devrait changer considérablement. De plus, lorsque nous devons nous prononcer sur la maîtrise d'un objectif, ces items devraient nous permettre de prendre des décisions

appropriées. Enfin, si les items en question proviennent d'un même domaine d'items, ils devraient être réussis ou échoués conjointement.

4.1 INDICE DE SENSIBILITÉ À L'ENSEIGNEMENT

Cox et Vargas (1966) ont proposé l'indice de sensibilité à l'enseignement pour déterminer quels items sont les plus affectés par l'enseignement. Cet indice est calculé en faisant la différence entre la difficulté d'un item après l'enseignement (p_{post}) et avant ($p_{pré}$) :

$$S = p_{post} - p_{pré} \quad (6.13)$$

Plus l'écart S est élevé, plus la mesure porte sur des items permettant de mesurer l'effet de l'enseignement. Moins S est élevé, moins l'item est utile car il a porté sur une question qui était tout aussi bien réussie avant l'enseignement qu'après. Un tel item ne permet pas de discriminer l'effet de l'enseignement.

Si au prétest un item est réussi par 23% des élèves et qu'au post-test il est réussi par 82%, la valeur de sensibilité à l'enseignement S = 0,82 - 0,23 = 0,59. Un tel résultat peut être interprété comme indiquant que l'item discrimine bien parmi les élèves qui ne réussissaient pas l'item avant l'enseignement et les élèves qui le réussissent maintenant.

Une valeur négative de S ou une valeur de 0 peuvent être interprétées de deux façons :

1. L'item ne convient pas, car il ne porte pas sur l'enseignement.
2. L'enseignement n'a eu aucun effet sur la réussite des élèves.

4.2 DISCRIMINATION AU SEUIL DE MAÎTRISE

Brennan (1972) a proposé un indice similaire à celui de Findley (1956) pour le calcul de la discrimination de l'item au seuil de maîtrise. Cet indice B est l'équivalent de l'indice D sauf que les groupes forts et faibles sont remplacés par les groupes qui ont atteint ou non le seuil de maîtrise au score total. L'indice B peut être calculé de la manière suivante :

$$B = p_M - p_{NM} \quad (6.14)$$

p_M représente l'indice de difficulté de l'item pour ceux qui ont atteint le seuil de maîtrise au test entier et p_{NM} représente l'indice de difficulté de l'item pour ceux qui ne l'ont pas atteint. B peut varier de -1 à +1. Un indice négatif signifie que l'item ne discrimine pas dans la même direction que les autres items au test. Un indice positif indique quelle proportion d'élèves dans le groupe « maîtrise » a mieux réussi l'item que dans le groupe « non maîtrise ».

Le tableau 5 présente un exemple de calcul de l'indice B. L'item de cette figure discrimine bien au seuil de maîtrise puisque, par rapport au groupe « non maîtrise », il y a 55% en plus d'élèves du groupe « maîtrise » qui le réussissent. C'est certainement un item adéquat pour différencier au seuil de maîtrise. Ce seuil est déterminé préalablement à l'examen. Un enseignant peut décider qu'un élève doit réussir 80% des

items d'un même domaine pour démontrer qu'il maîtrise un objectif. L'élève qui obtient 80% et plus au test sera considéré comme appartenant au groupe « maîtrise », alors que les autres élèves (moins de 80%) feront partie du groupe « non maîtrise ».

$$p_M = \frac{b}{b+d} = \frac{8}{8+2} = 0{,}8 \qquad (6.15)$$

$$p_{NM} = \frac{a}{a+c} = \frac{4}{4+12} = 0{,}25 \qquad (6.16)$$

$$B = p_M - p_{NM} = 0{,}8 - 0{,}25 = 0{,}55 \qquad (6.17)$$

Tableau 5 – Exemple de calcul du coefficient de Brennam

Item	*Réussi* a + b	a = 4	b = 8
	Échoué c + d	c = 12	d = 2
		Non maîtrise a + c	*Maîtrise* b + d
		Test	

4.3 ÉQUIVALENCE DES ITEMS APPARTENANT À UN MÊME DOMAINE

La préparation d'instruments de mesure critériée nous amène à construire des items faisant partie d'un même domaine. L'analyse des items devrait nous permettre de vérifier a posteriori si tel est bien le cas. Des items appartenant à un même domaine devraient être réussis ou échoués conjointement, ce qui devrait se traduire par un manque d'indépendance dans la distribution conjointe de ces deux items. Un test du χ^2 permet de vérifier si la distribution des fréquences conjointes est significativement différente de celle à laquelle on pourrait s'attendre si une telle distribution s'était produite aléatoirement. Le tableau 6 présente un exemple de données servant au calcul du χ^2 entre deux items.

Il y a deux façons de calculer la valeur du χ^2 au tableau 6. La première, plus générale, s'applique à toutes les situations. Elle nécessite le calcul de fréquences théoriques *FT* (inscrites entre parenthèses dans chaque cellule du tableau) qui se produiraient s'il n'y avait aucune association entre les deux items. Plus les fréquences observées *FO* (a, b, c, d) sont différentes des fréquences théoriques *FT*, plus il est

permis de croire que les items ne sont pas indépendants entre eux mais qu'ils sont associés et mesurent le même domaine.

Tableau 6 – Association entre deux items A et B

		Item B: *Échoué*	Item B: *Réussi*	
Item A	*Réussi*	a = 6 (FT = 8)	b = 14 (FT = 12)	a + b = 20
	Échoué	c = 6 (FT = 4)	d = 4 (FT = 6)	c + d = 12
		a + c = 12	b + d = 18	

Le calcul des fréquences théoriques est fort simple. Il s'agit de trouver, pour chaque cellule du tableau de contingence, la fréquence qui respecte les proportions des totaux marginaux. Ainsi, si 20 élèves sur 30 ont réussi l'item A et que 18 élèves sur 30 ont réussi l'item B, alors 20/30 des 18 élèves de l'item B devraient réussir conjointement les items A et B, soit 12 élèves. Les autres fréquences théoriques se déduisent par soustraction. Pour trouver les fréquences théoriques de la cellule a, il suffit de soustraire la fréquence théorique 12 du total marginal de cette rangée : 20 - 12 = 8. La somme des fréquences théoriques pour chaque rangée et chaque colonne doit correspondre aux totaux marginaux.

Une fois élevée au carré et divisée par la fréquence théorique, la somme des écarts entre fréquences théoriques et fréquences observées nous donne la valeur du χ^2. L'équation (6.18) résume le calcul du χ^2 selon cette méthode.

$$\chi^2 = \sum \frac{(FT - FO)^2}{FT} \tag{6.18}$$

Dans le cas de l'exemple du tableau 6, la valeur calculée du χ^2 serait :

$$\chi^2 = \sum \frac{(8-6)^2}{8} + \frac{(12-14)^2}{12} + \frac{(6-4)^2}{4} + \frac{(4-6)^2}{6} = 2{,}5 \tag{6.19}$$

Lorsque chaque item ne peut prendre que deux valeurs, l'équation (6.18) peut être remplacée par l'équation (6.20) où n'interviennent que les fréquences observées :

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)\,(c+d)\,(b+d)\,(a+c)} \tag{6.20}$$

Chaque lettre correspond aux cellules du tableau 6. En substituant chaque lettre par la fréquence observée correspondante, on retrouve la même valeur de χ^2 calculée par l'équation (6.18). En effet,

$$\chi^2 = \frac{30\,(24-64)^2}{20 \times 10 \times 12 \times 18} = 2{,}5 \tag{6.21}$$

Comment interpréter la valeur calculée du χ^2 ? Pour cela, il est nécessaire de consulter une table des probabilités des valeurs du χ^2 pour un degré de liberté. La valeur critique pour α=0,05 étant de 3,84, nous savons qu'une valeur de 2,5 a plus de 5 chances sur 100 de se produire au hasard. Comme la valeur obtenue est inférieure à la valeur critique, nous pouvons considérer ces deux items comme indépendants, donc sans association entre l'un et l'autre. Il serait donc difficile de considérer ces deux items comme provenant du même domaine.

Lorsque le nombre de catégories de chaque item excède deux, l'équation (6.20) ne permet pas de calculer la valeur du χ^2. Il faut alors avoir recours à l'équation (6.18). Le nombre de degré de liberté est égal à *[(c-1) (r-1)]*, où *r* et *c* représentent le nombre de catégories de l'item A (rangées) et de l'item B (colonnes).

En règle générale, il est préférable d'employer l'équation (6.18). D'abord, parce qu'elle permet de découvrir dans quelle(s) cellule(s) les différences entres fréquences observées et fréquences théoriques sont les plus grandes. Ensuite, parce que la valeur du χ^2 est biaisée lorsque plus de 20% des fréquences théoriques sont inférieures à 5 ou encore lorsqu'elles sont inférieures à 1 ou égales à 0.

Il n'est pas toujours nécessaire de calculer un χ^2 pour se faire une opinion à propos du degré de concordance entre deux items A et B. Harris et Pearlman (1977) ont proposé de calculer une proportion d'accord, tel que *(b+c)/n*. C'est un moyen simple de calculer quelle proportion d'élèves ont fourni le même résultat aux deux items. Dans le cas de l'exemple du tableau 6, la proportion d'accord est de 20/30, soit 0,67. Cette proportion signifie que 33% des élèves ont réussi un item sans avoir réussi l'autre. Il s'agit d'une proportion suffisamment élevée pour ne pas considérer les deux items comme provenant du même domaine.

Harris et Pearlman (1977) ont également proposé un moyen de vérifier si deux items sont de même difficulté. En effet, si deux items sont rédigés à partir du même objectif d'apprentissage et qu'ils ont fait l'objet en classe d'efforts de préparation comparables, ceux-ci devraient être de difficultés sensiblement identiques. Un test de signification sur la différence entre la difficulté de deux items devrait nous permettre de décider si un écart est suffisamment grand pour considérer les items comme appartenant à deux domaines différents ou s'il est possible que la différence observée soit purement fortuite.

Pour tester ces possibilités, Harris et Pearlman proposent le test du χ^2 suivant avec un degré de liberté :

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \tag{6.22}$$

Il est possible d'appliquer la procédure proposée par Harris et Pearlman (1977) à l'exemple du tableau 5. Les items A et B ont respectivement 20/30 et 18/30 comme indices de difficulté. Une différence de 2/30 est-elle suffisante pour considérer que les deux items ont des degrés de difficulté différents ou cet écart peut-il être attribué aux

effets de l'échantillonnage ? Pour répondre à cette question, calculons la valeur de selon l'équation (6.23) en substituant b et c par leurs valeurs respectives. Ceci nous fournit le résultat suivant :

$$\chi^2 = \frac{(|14-6|-1)^2}{14+6} = 2{,}45 \qquad (6.23)$$

La valeur calculée est inférieure à la valeur critique de 3,81 pour un niveau de signification $\alpha = 0{,}05$. Il faut donc considérer que les items A et B ne sont pas de degrés de difficulté significativement différents, puisqu'il y a plus de cinq chances sur 100 qu'un écart de 2/30 entre les deux items soit dû aux effets d'échantillonnage.

Il peut sembler paradoxal que deux items que nous avons déclaré comme appartenant à des domaines différents possèdent des degrés de difficulté équivalents. En fait, si deux items appartiennent au même domaine, ils seront nécessairement de même degré de difficulté. Par contre, deux items de domaines différents, tels que les items A et B du tableau 5 peuvent être de degrés de difficulté semblables. Même en appartenant à des domaines différents, rien n'empêche qu'ils puissent être réussis par des proportions égales de sujets. Ce serait le cas, par exemple, d'un item de géographie et d'un item de français réussis par 12 élèves sur 24 (50%).

En guise de conclusion, soulignons que ce dernier test de Harris et Pearlman ne nous permet pas de nous prononcer quant à savoir si deux items appartiennent au même domaine. En effet, comme nous venons de le voir, l'absence de différences significatives des degrés de difficulté de deux items constitue une condition nécessaire mais non suffisante à ce qu'ils appartiennent au même domaine.

5. Les indices de fiabilité et de validité

En plus des indices de difficulté et de discrimination, il existe deux autres indices fort utiles lors d'une analyse d'items : l'*indice de fiabilité* et l'*indice de validité*. La contribution respective de chaque item à la fiabilité et à la validité du test entier peut nous aider à optimiser notre instrument de mesure en ne choisissant que les items les plus pertinents à nos objectifs d'évaluation.

Ces indices nous sont donnés par la corrélation item-total de chaque item pondérée par son écart type. La corrélation item-total est calculée soit avec un critère interne (X = score total au test), soit avec un critère externe (Y = score total au critère). Dans le premier cas, nous obtenons l'indice de fiabilité. Dans le second cas, il s'agit de l'indice de validité.

L'indice de fiabilité est donc fourni par le produit, où s_i est l'écart type de l'item et r_{iX} est la corrélation item-total. L'indice de validité se calcule de la même manière par le produit $s_i\ r_{iY}$. Dans ce dernier cas, r_{iY} représente la corrélation item-critère.

5.1 ANALYSE DES ITEMS À PARTIR DES INDICES DE FIABILITÉ ET DE VALIDITÉ

La figure 6 présente la forme que pourrait prendre une analyse d'items visant à optimiser la validité et la fiabilité d'un test à partir de ces indices. En situant chaque item en fonction de son indice de fiabilité et de son indice de validité dans un plan cartésien, il devient relativement simple de choisir les items qui contribuent à accroître simultanément la fiabilité du score total au test et sa validité par rapport au critère choisi. Les items 1 à 6 présentent à cet égard six situations caractéristiques :

1. L'item 1 est l'item idéal. Il possède des indices élevés de fiabilité et de validité. C'est certainement le genre d'item que nous souhaiterions conserver.
2. L'item 2 est un item fidèle, mais de validité moyenne. Il contribue à la précision du test, mais peu à sa pertinence par rapport au critère.
3. L'item 3 est un item sans validité. Si la fiabilité du test était notre seul souci, on pourrait opter pour le conserver. Mais à quoi sert-il de conserver un item qui n'est pas valide ? Ce n'est pas l'item à privilégier si nous cherchons à accroître la validité de notre test.
4. L'item 4 est un item ayant une meilleure relation avec le critère qu'avec le score total au test. C'est donc un item qui mesure une caractéristique importante dans la prédiction du critère qui n'est pas mesurée par le test actuel. Il faudrait considérer si un nouveau test constitué de ce genre d'items ne constituerait pas un test plus valide que le test actuel. L'autre solution serait de créer deux sous-tests, chacun mesurant des caractéristiques différentes et indépendantes du critère.
5. L'item 5 est un item semblable à l'item 4, sauf qu'il est un peu moins valide.
6. L'item 6 est le prototype des items qui ne sont d'aucune utilité, que ce soit sur le plan de la validité des résultats ou de la fiabilité. Tous les items de la zone grise devraient être rejetés ou révisés en profondeur. En effet, ils constituent une perte de temps puisqu'ils contribuent très peu à la précision et à la pertinence du test.

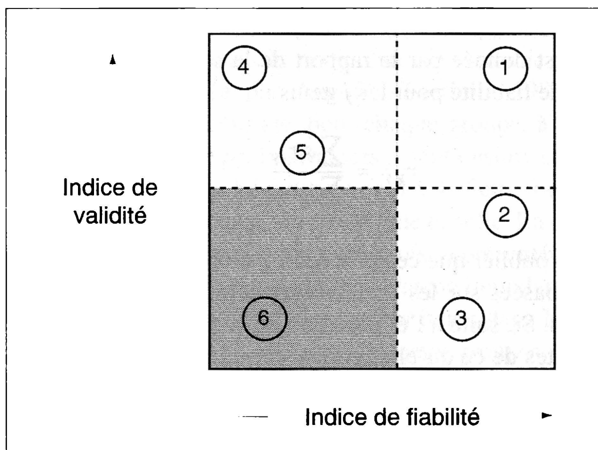


Figure 6 – Sélection des items en fonction des indices de fiabilité et de validité

5.2 Optimisation d'un test

En préparant un nouveau test, il est préférable de rédiger un nombre d'items plus grand que ce que nous prévoyons utiliser. Ceci permettra d'élaborer un meilleur test en ne choisissant que les items qui auront les caractéristiques souhaitées : c'est ce que nous appelons l'*optimisation* des caractéristiques d'un test.

À cet égard, les indices de fiabilité et de validité possèdent des propriétés intéressantes qu'il nous est possible d'exploiter lors d'une étude d'optimisation. Par exemple, on peut démontrer que la variance des scores totaux à un test est égale à la somme des indices de fiabilité.

$$s_X^2 = \left(\sum s_i r_{iX}\right)^2 \tag{6.24}$$

Le concepteur d'un test peut donc choisir d'additionner les indices de fiabilité des items jusqu'à ce qu'il obtienne la variance des résultats souhaitée. Il débutera par les items dont les indices de fiabilité sont les plus élevés jusqu'à ce qu'il ait atteint la variance souhaitée avec le minimum d'items nécessaires.

Le même exercice peut être répété en ce qui concerne la fiabilité de cohérence interne. En effet, on peut démontrer que le coefficient α peut également s'exprimer en fonction des indices de fiabilité. L'équation (6.24) exprime cette relation de la manière suivante :

$$\alpha = \frac{j}{j-1}\left[1 - \frac{\sum s_i^2}{\left(\sum s_i r_{iX}\right)^2}\right] \tag{6.25}$$

Au dénominateur de l'équation (6.25), on reconnaît l'expression de la variance totale du test exprimée en fonction de la somme des indices de fiabilité des items (voir équation 6.24). Cette équation permet donc d'estimer la fiabilité qu'aurait un test constitué des j items sélectionnés.

La même procédure peut également servir à calculer la validité à partir des j meilleurs items. Dans ce cas, la validité du nouveau score total formé de la somme de j items sélectionnés est donnée par le rapport de la somme des indices de validité à la somme des indices de fiabilité pour les j items :

$$r_{XY} = \frac{\sum s_i r_{iY}}{\sum s_i r_{iX}} \tag{6.26}$$

Il ne faut pas oublier que ces procédures d'optimisation ne sont qu'approximatives car elles sont basées sur les corrélations item-total et item-critère calculées sur *l'ensemble* des items. Si, suite à l'élimination de certains items, ces corrélations devaient être fort différentes de ce qu'elles étaient initialement, les valeurs de variance, fiabilité et validité calculées par les équations (6.24 à 6.26) pourraient être différentes de celles que l'on aurait obtenues en refaisant les calculs à partir des nouveaux scores totaux. Pour en savoir plus sur le développement de ces équations, on peut consulter Gulliksen (1950) et Lord et Novick (1968).

6. Le fonctionnement différentiel des items

Lors de la construction d'un test, une attention particulière doit être portée à la validité différentielle du contenu de ce test pour les différents sous-groupes qui composent la population à laquelle il est destiné. Cette validité de contenu peut être évaluée de deux manières. La première s'appuie sur l'évaluation de chaque item par un groupe d'experts. La seconde est mathématique et utilise des techniques statistiques appliquées aux données recueillies pour les items étudiés.

Différentes recherches ont montré (Hambleton & Jones, 1993) que les méthodes statistiques d'analyse de contenu sont les plus objectives et les plus efficaces pour repérer les items biaisés au sein d'un test. Dans ce type d'analyse, *« un item est considéré comme non biaisé lorsque la probabilité de réussir cet item est la même pour tous les sujets de la population possédant la même aptitude, indépendamment de leur sous-groupe d'appartenance »* (Osterlind, 1989, p.11). Ainsi, il est erroné de croire qu'un item est biaisé uniquement parce qu'il existe une différence de performance entre deux groupes. Pour qu'il y ait biais, il est nécessaire que les sujets des deux groupes se situent au même niveau d'aptitude. Vu le caractère général de la notion de biais et les problèmes d'interprétation qui en résultent, certains auteurs (par exemple, Holland & Thayer, 1988, p.129) ont proposé de préférer à ce terme l'expression *« fonctionnement différentiel de l'item »* (*« differential item functioning »*, en abrégé DIF). À présent, l'usage de ce dernier terme a largement supplanté celui de biais dans la littérature consacrée à l'analyse des items. Nous l'utiliserons donc dans la suite de cette section.

Les méthodes statistiques permettant d'analyser le fonctionnement différentiel des items peuvent être regroupées en deux grandes catégories (Scheuneman & Bleinstein, 1989) : (1) les méthodes basées sur les résultats observés aux items et sur le score au test et qui se réfèrent au modèle classique de la mesure ; (2) les méthodes basées sur les aptitudes « vraies » et qui se réfèrent aux modèles de la réponse à l'item (voir chapitre 8).

Du fait de leur simplicité théorique et de leur facilité pratique, les méthodes appartenant au premier groupe ont été les premières développées et appliquées. Parmi celles-ci, une méthode a connu un succès particulier : la *méthode du graphique Delta (« delta-plot method »)* développée par Angoff (Osterlind, 1989 ; Scheuneman & Bleinstein, 1989). Cette méthode consiste, pour chaque groupe, à calculer l'indice de difficulté de chaque item (sa valeur *p*). Les valeurs *p* sont ensuite converties en valeurs Δ dont la moyenne est égale à 13 et l'écart type égal à 4. Cette transformation permet de placer les valeurs *p* des deux groupes sur une même échelle. La distribution bivariée de la difficulté des items ainsi transformés est alors représentée sur un graphique (figure 7). Sur l'abscisse, on reporte la valeur Δ de chaque item dans le premier groupe et, sur l'ordonnée, on reporte la valeur Δ des mêmes items dans le second groupe. Si chaque item présente une même difficulté dans les deux groupes, la relation entre les valeurs Δ de ces deux groupes prend la forme d'une droite. Mais, le plus souvent, cette relation a la forme d'un nuage de points. Les items présentant un fonctionnement différentiel sont ceux qui s'écartent le plus de la droite de régression qui peut être tracée au sein du nuage de points.

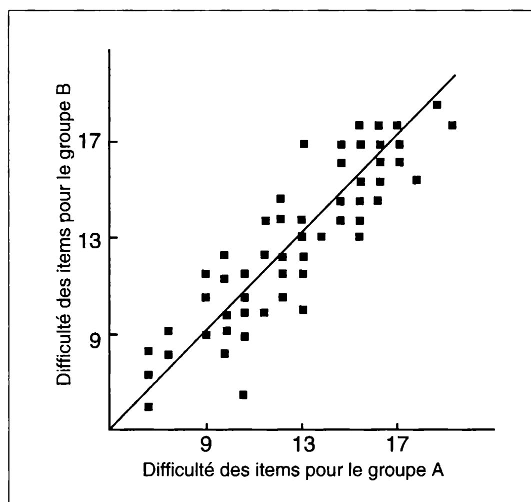


Figure 7 – Exemple de distribution bivariée de la difficulté des items d'un test

La méthode du graphique Delta a été vivement critiquée (Cole & Moss, 1989, p.209) car elle postule que les sujets des deux groupes possèdent le même niveau d'aptitude. Or, le plus souvent, ce n'est pas le cas. Comme la méthode du graphique Delta ne prend pas en compte les propriétés de discrimination des items, cela conduit à considérer erronément certains items comme fonctionnant de manière différentielle. Pour éviter ce type de problème, il est nécessaire d'utiliser des méthodes conditionnelles, c'est-à-dire des méthodes qui comparent la difficulté des items uniquement entre sujets de même niveau d'aptitude. Certaines de ces méthodes, comme celle de Mantel-Haenszel présentée ci-dessous, se réfèrent à la théorie classique des tests. D'autres s'appuient sur les modèles de réponse à l'item (MRI). Nous parlerons plus en détail de ces dernières dans le chapitre 8, consacré aux MRI.

La méthode de Mantel-Haenszel a été largement adoptée par les psychométriciens car elle a l'avantage d'être simple à utiliser et de permettre un bon repérage des items problématiques à partir d'échantillons de taille moyenne (N=200 par groupe). Par ailleurs, elle comprend un test de signification ainsi qu'un indice permettant d'apprécier l'importance du fonctionnement différentiel.

Cette méthode a été développée, il y a plus de 30 ans, dans le domaine médical par Mantel et Haenszel. Mais elle n'a été utilisée que récemment en psychométrie sous l'impulsion de Holland et Thayer (1988) qui ont démontré son intérêt pour l'analyse du fonctionnement différentiel des items.

La méthode de Mantel-Haenszel consiste à comparer la chance de réussir un item pour les membres de deux groupes après que les individus aient été pairés sur base d'une aptitude déterminée. Le groupe dont on veut étudier les résultats aux items est habituellement appelé le groupe focal (F). Le groupe dont les performances sont prises comme base de comparaison est appelé le groupe de référence (R). La première étape de l'analyse consiste à déterminer le niveau d'aptitude de chaque individu au sein des deux groupes. L'aptitude en question est celle mesurée par le test dont on étudie les items. Elle peut être évaluée à l'aide d'un critère externe comme, par exemple,

le résultat à un autre test. Mais, le plus souvent, le score total au test étudié est pris comme critère interne de classement des individus. Une fois les sujets rangés en catégories, il est alors possible de pairer celles-ci entre le groupe focal et le groupe de référence.

Pour chaque catégorie, une table de contingence 2 x 2 peut alors être construite. Cette table compare la fréquence de réussite et d'échec d'un item dans le groupe focal et dans le groupe de référence. Pour chaque item du test, il y a autant de tables de contingence 2 x 2 que de catégories d'aptitude. Le tableau 6 (d'après Holland & Thayer, 1988, p.130) illustre la forme générale de chaque table de contingence. Dans ce tableau, T_j représente le nombre total de sujets d'un niveau d'aptitude donné, n_{Rj} représente le nombre de sujets du groupe R et A_j représente le nombre de sujets du groupe R qui ont réussi l'item. Les autres entrées du tableau se définissent de manière similaire.

Tableau 7 – Table de contingence pour la jème catégorie pairée de sujets des groupes R et F

		Résultat à l'item		
		1	0	Total
Groupe	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	n_{Fj}
	Total	m_{1j}	m_{0j}	T_j

L'hypothèse selon laquelle un item ne présente pas de fonctionnement différentiel correspond à l'hypothèse nulle. Dans ce cas, pour tous les niveaux j d'aptitude, le groupe focal et le groupe de référence ont des performances identiques à l'item en question. Cette hypothèse peut être testée au moyen du test χ^2 de Mantel-Haenszel (χ^2_{m-h}). Sous l'hypothèse nulle, χ^2_{m-h} se distribue approximativement comme χ^2 avec un degré de liberté.

$$\chi^2_{m-h} = \frac{(|\sum A_j - \sum E(A_j)| - 0{,}5)^2}{\sum s^2(A_j)} \qquad (6.27)$$

Dans cette équation,

$$E(A_j) = \frac{n_{Rj} m_{1j}}{T_j} = \text{valeur attendue de } A_j \qquad (6.28)$$

$$s^2(A_j) = \frac{n_{Rj}n_{Fj}m_{1j}m_{0j}}{T_j^2(T_j - 1)} = \text{variance de } A_j \tag{6.29}$$

Outre un test de signification, la procédure proposée par Mantel et Haenszel inclut une estimation du rapport des résultats entre les deux groupes. Cette estimation est donnée par l'équation suivante :

$$\hat{a} = \sum \frac{A_j D_j}{T_j} / \sum \frac{BjC_j}{T_j} \tag{6.30}$$

La valeur de α peut varier de 0 à ∞. Une valeur égale à 1 signifie qu'il n'y a pas de fonctionnement différentiel. Vu sa distribution asymétrique, ce coefficient est peu aisé à interpréter. Pour cette raison, on préfère utiliser le logarithme de $\hat{\alpha}$ qui permet d'obtenir un index, appelé Delta (Δ), qui se distribue symétriquement autour de 0 qui est la valeur nulle :

$$\Delta = 2{,}35 \ln(\hat{\alpha}) \tag{6.31}$$

La valeur absolue de Δ représente la différence du niveau moyen de difficulté entre les deux groupes. Le signe de Δ indique la direction de cette différence. Une valeur positive indique que l'item est relativement plus facile pour le groupe focal. Inversement, une valeur négative indique que l'item est relativement plus facile pour le groupe de référence. Selon Dorans (1989), il faut considérer qu'un item présente un fonctionnement différentiel important lorsque la valeur du test χ^2_{m-h} est significative et que la valeur absolue de Δ est égale ou supérieure à $|1{,}50|$.

Nous avons souligné plus haut les nombreux avantages de la méthode de Mantel-Haenszel. Elle présente cependant certaines limites. La première concerne le critère permettant de définir et ensuite de pairer les niveaux d'aptitude des sujets des deux groupes. Nous avons souligné que, le plus souvent, le critère utilisé est le score total au test lui-même. L'usage d'un critère interne ne va pas sans problème. En effet, les items qui présentent un fonctionnement différentiel important interviennent dans le score total et peuvent donc le fausser. Pour éviter ce problème, il est d'usage (Hambleton & al., 1993) d'appliquer la méthode de Mantel-Haenszel en deux étapes. Lors de la première étape, tous les items interviennent dans le score total. Les items repérés comme présentant un fonctionnement différentiel sont alors exclus du score total et une seconde analyse est réalisée. Lors de cette seconde étape, l'usage d'un score total « épuré » permet un repérage mieux assuré des items problématiques. Il semble que cette manière de faire soit préférable à l'usage d'un critère externe dont l'adéquation et la fiabilité risquent d'être moins bonnes (Angoff, 1993).

Un autre problème, lié au choix du critère, est celui du nombre de catégories au sein desquelles regrouper les sujets. Holland et Thayer (1988) recommandent d'utiliser *k+1* catégories ; *k* étant le nombre d'items du test. Différentes recherches (Hambleton & al., 1993) ont montré que la réduction du nombre de catégories n'améliorait que faiblement la puissance du test statistique lorsque la distribution de l'aptitude était équivalente dans les deux groupes. Par contre, lorsque la distribution de l'aptitude est inégale dans les deux groupes, la réduction du nombre de catégories améliore la détection des items problématiques, mais au prix d'une augmentation de l'erreur de type I.

Par conséquent, afin d'améliorer la puissance du test statistique, il est préférable d'augmenter la taille des échantillons plutôt que de réduire le nombre de groupes de sujets.

Un autre problème concerne la taille des échantillons. Nous avons souligné qu'un des avantages de la méthode de Mantel-Haenszel est de ne nécessiter que des échantillons de taille moyenne. Plusieurs recherches (Mazor & al., 1992) indiquent que la taille minimum de chaque échantillon devrait être d'environ 200 sujets. En dessous de cette taille, le nombre d'items problématiques non repérés augmente notablement. D'une manière générale, plus la taille des échantillons est grande, plus sensible est l'évaluation des items. Toutefois, si l'objectif est de repérer les items les plus problématiques, la méthode de Mantel-Haenszel reste la méthode de choix lorsque l'on ne dispose que d'échantillons relativement réduits.

Enfin, un dernier problème posé par la méthode de Mantel-Haenszel concerne la détection du fonctionnement différentiel non uniforme de certains items. Il s'agit d'items dont le sens de la différence entre groupes varie selon le niveau d'aptitude des sujets. Par exemple, lorsque les sujets possèdent un faible niveau d'aptitude, un item sera plus difficile dans le groupe F que dans le groupe R. Par contre, lorsque les sujets possèdent un haut niveau d'aptitude, nous observerons le phénomène inverse (figure 8). On constate que la méthode de Mantel-Haenszel détecte mal de tels items (Hambleton & Rogers, 1989). Dans ce cas, le recours à d'autres méthodes comme celles basées sur la comparaison des courbes caractéristiques d'items s'impose.

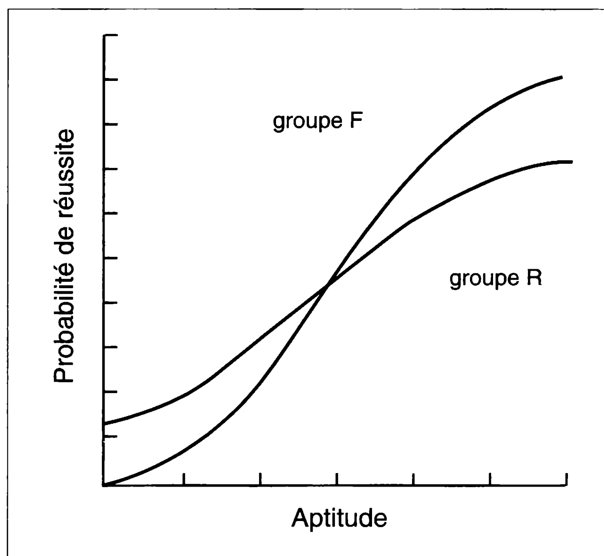


Figure 8 – Courbes caractéristiques d'un même item pour deux groupes dans le cas d'un fonctionnement différentiel non uniforme

Dans le cas de tests construits à partir de vastes banques d'items, certains chercheurs choisissent d'éliminer de manière systématique tous les items repérés comme problématiques à la suite des analyses statistiques. Mais, dans la majorité des cas, une telle politique n'est économiquement pas possible. Il est en effet difficile de créer et de pré-tester à grande échelle plus d'un certain nombre d'items. Il est dès lors nécessaire

d'analyser le cas de chaque item repéré et de voir s'il n'est pas possible de le conserver malgré de mauvais indices statistiques. Plusieurs règles doivent être suivies lors de cette interprétation (Nardakumar & al., 1993). La première est de tenir compte du risque d'erreur de type I (rejeter erronément H_0) dû au grand nombre de tests statistiques réalisés de manière simultanée (Dechef & Laveault, 1993). Par exemple, si nous testons au même moment le fonctionnement différentiel de 150 items, nous devons nous attendre à observer un certain nombre de χ^2 significatifs alors que H_o est vraie (absence de fonctionnement différentiel entre les deux groupes).

Un autre principe à prendre en compte lors de l'interprétation des items problématiques est leur poids effectif au sein du test. Un seul item présentant un léger fonctionnement différentiel dans un test de 30 items n'est pas véritablement un problème. Un principe complémentaire de celui-ci est d'être attentif à l'équilibre entre les items problématiques dans les deux groupes étudiés. Si, par exemple, un test comprend deux items qui favorisent les garçons et deux items qui favorisent les filles, le résultat global à ce test ne sera pas affecté par le fonctionnement différentiel des items qui le composent. Les différents items problématiques se contrebalancent en effet au niveau du score total (Grégoire, 1995, pour une illustration).

7. Choisir l'analyse d'items appropriée au type d'évaluation

Les techniques d'analyse d'items sont nombreuses et variées. Chacune poursuit un but précis et nous fournit une information précieuse sur le rôle joué par chaque item dans le score total. L'analyse d'items est donc indispensable à toute opération de mesure qui se veut valide et fiable.

Chaque type d'évaluation fait appel à des techniques particulières d'analyse d'items. C'est ce que résume le tableau 7. Si le but de l'évaluation est de discriminer parmi les sujets, comme c'est le cas en psychologie avec les tests de sélection du personnel ou en éducation lorsqu'il s'agit d'évaluation sommative, l'analyse d'items va privilégier les items qui discriminent fortement les sujets, de même que les items dont les indices de validité et de fidélité sont élevés. Enfin, une analyse démontrant un fonctionnement différentiel de certains items pourra contribuer à éliminer ceux qui mesurent une caractéristique sans rapport avec le trait mesuré, contribuant ainsi à biaiser les résultats en faveur d'un groupe ou d'un autre.

Tableau 8 – Techniques d'analyses d'items selon le type d'évaluation

	Indices de difficulté	Indices de discrimination	Indices de fiabilité et de validité	Indicateurs de biais
Évaluation sommative, épreuves de sélection	p, p_c	D	$r_{iX}\, r_{iY}$	χ^2_{m-h}
Évaluation formative (mesure critériée, tests de maîtrise)	p	B	χ^2	—

La plupart de ces indicateurs ne sont guère utiles en évaluation formative, que celle-ci repose sur des épreuves de maîtrise ou des instruments de mesure critériée. La différenciation des sujets n'est pas importante et la validité et la fiabilité, quoique toujours importantes, ne donnent pas lieu à des analyses poussées puisque l'évaluation ne va pas aboutir à une décision finale quant au classement du sujet. Le but de l'évaluation formative est plutôt d'aider et de remédier à une situation qui comporte des difficultés pour le sujet. Pour les mêmes raisons, l'étude du fonctionnement différentiel des items intéresse fort peu l'évaluation formative.

L'analyse d'items en évaluation formative se limite aux approches formelles instrumentées, telles que celle de la pédagogie de la réussite ou encore celle de la mesure critériée. Dans le cas de tests de maîtrise, la discrimination la plus importante se situe au seuil de réussite. L'indice de Brennan est particulièrement utile dans ce contexte. Dans le cas de tests critériés, les tests du khi-deux permettent de vérifier si les habiletés mesurées sont de difficultés comparables ou si elles proviennent du même domaine. Ce genre de vérification peut être utile surtout si l'on songe à regrouper les réussites à certaines catégories d'items pour constituer non pas un score total, mais un *profil de scores*.

L'ensemble des techniques précédentes permet d'analyser les propriétés des items en rapport avec les valeurs des scores observés des sujets. Ces techniques conviennent particulièrement en éducation et en psychologie lorsque les échantillons sont petits. Il faut toutefois se rappeler que la valeur des conclusions de ces analyses se limite aux échantillons étudiés et aux populations dont ils sont tirés.

Lorsque l'on souhaite faire porter l'analyse sur les caractéristiques sous-jacentes aux items (traits latents), les analyses d'items permises par les modèles des réponses aux items sont beaucoup plus puissantes en autant que nous disposions d'échantillons de sujets et d'items de grandes tailles. C'est pourquoi ces analyses, décrites dans le chapitre 8, conviennent particulièrement aux opérations de testing à grande échelle telles que les enquêtes nationales ou internationales.

CHAPITRE 7
CALCUL ET INTERPRÉTATION DES SCORES

1. Les normes

1.1 ÉCHELLES NORMÉES ET NON NORMÉES

Dans le cadre d'une évaluation normée, tester consiste toujours à comparer des sujets, à les distinguer entre eux. Sans référence aux résultats d'autres sujets, les notes brutes d'un individu à un test donné sont sans signification précise. En effet, d'un test à l'autre, la nature et la difficulté des items varient. Sur base d'un score brut, nous ne pouvons donc déterminer si un sujet est faible ou brillant. Pour pouvoir interpréter les résultats, il est nécessaire de faire correspondre les notes brutes à celles d'une échelle qui possède une signification normative. Nous verrons plus loin que cette échelle peut prendre différentes formes plus ou moins commodes pour le praticien.

L'étalonnage d'un test est la graduation des résultats de celui-ci, la fixation d'échelons qui vont permettre la comparaison des résultats de divers individus. Pour étalonner un test, celui-ci doit être passé par un échantillon représentatif de la population choisie. Les résultats obtenus serviront alors de normes de référence pour cette population et elle seule. Des écarts parfois sensibles existent non seulement entre les performances de populations dont la culture et le système éducatif sont très différents mais aussi entre des populations plus proches en apparence.

La relativité des normes n'est pas seulement synchronique mais elle est aussi diachronique. Les caractéristiques d'une population ne restent en effet pas stables au cours du temps. La composition d'une population peut changer et, surtout, les conditions éducatives peuvent se modifier. Par exemple, dans les pays occidentaux, on observe que l'élévation du niveau moyen de scolarité entraîne une augmentation des performances aux tests d'intelligence. C'est ce que constate J.R. Flynn (1987) dans une importante recherche internationale. Entre autres, l'auteur rapporte des données très fiables à propos des performances des appelés hollandais qui, chaque année, for-

ment un important échantillon d'hommes âgés de 18 ans. Entre 1952 et 1982, tous ces appelés ont passé le même test d'intelligence, en l'occurrence les Matrices de Raven. Si nous prenons comme normes de référence celles de 1952, nous constatons que le QI moyen de l'échantillon de 1982 atteint 121,10 points. En trente ans, nous observons ainsi un bond de plus de vingt points de QI à un test qui, rappelons-le, est non verbal. Des données similaires ont été recueillies sur les appelés belges à l'aide du même test de Raven, mais l'étude a été faite sur une période plus brève (de 1958 à 1967). Sur cette période, on observe une augmentation du Q.I. de 6,47 points chez les belges francophones et de 7,82 points chez les belges néerlandophones. En France, toujours chez les appelés et toujours avec le test de Raven, le saut quantitatif est encore plus spectaculaire puisqu'entre 1949 et 1974, le Q.I. moyen a augmenté de 25,12 points.

Le praticien doit donc garder en tête que les normes vieillissent. La vitesse de cette dégradation de la qualité des normes varie toutefois selon le type de test. Les normes d'un test de développement psychomoteur bougent peu avec le temps. Par contre, un test de vocabulaire ou un questionnaire de personnalité voient leurs normes changer beaucoup plus rapidement. Un réétalonnage régulier des tests est donc une nécessité.

Angoff (1971), constatant que les échelles qui possèdent une signification normée sont condamnées à devenir obsolètes avec le temps, souligne l'intérêt de créer des échelles non normées, c'est-à-dire indépendantes de tout groupe de sujets. Les échelles construites dans le cadre de l'évaluation critériée sont un exemple d'échelles non normées. Mais, c'est surtout dans le cadre des modèles de la réponse à l'item que des échelles non normées ont pu être développées. Dans ce cas, la difficulté d'un item est considérée comme un paramètre invariant, indépendant de l'échantillon de sujets qui a permis de l'estimer. Avec un ensemble d'items, il est dès lors possible de construire une échelle de référence sans caractère normé. Cette question, loin d'être triviale, sera traitée plus en détail dans le chapitre 8 consacré aux modèles de la réponse à l'item.

1.2 ÉTABLISSEMENT DES NORMES

1.2.1 *Définition de la population*

Comme nous l'avons indiqué plus haut, la procédure d'étalonnage d'un test comprend la passation de celui-ci par un échantillon de la population de référence. Il est donc nécessaire de débuter la procédure par une définition claire de cette population. Rappelons que, du point de vue statistique, une population est l'ensemble de tous les cas qui font l'objet de l'intérêt du chercheur. Cet ensemble peut être fini ou infini. La population peut parfois être constituée d'un petit nombre de cas qui peuvent être tous mesurés. Mais, le plus souvent, la taille de la population rend toute collecte exhaustive difficile, voire impossible. Il faut alors se limiter à un échantillon à partir duquel les caractéristiques de la population seront inférées.

La définition de la population doit être appropriée à l'usage qui sera fait du test. Par exemple, si un test est destiné à diagnostiquer les troubles du développement sensori-moteur, la population visée sera celle des enfants âgés de 0 à 2 ans et demi. Et si un questionnaire est destiné à évaluer le développement social des handicapés mentaux, la population de référence sera celle des handicapés mentaux. D'une manière générale, il est nécessaire que la population de référence soit suffisamment homogène,

c'est-à-dire que tous les individus susceptibles d'être évalués à l'aide du test en fassent clairement partie.

Lorsqu'un test est développé par un éditeur commercial, il est fréquent que les normes soient nationales. L'avantage majeur de se référer à une population nationale est de permettre la production d'un système unique de normes, valable pour un très grand nombre de sujets. L'intérêt commercial et la facilité d'usage sont évidents. Cependant, la référence à la population nationale n'implique pas que les normes des différents tests soient ipso facto comparables. En effet, cette population n'est pas toujours définie de la même manière par les éditeurs. En particulier, ces derniers ne s'accordent pas à propos de l'inclusion de certains groupes atypiques dans la population de référence. Par exemple, les handicapés mentaux sont parfois inclus et d'autres fois exclus de la population. Il en résulte des différences sensibles entre les normes de certains tests qui, pourtant, se réfèrent tous à la population nationale. Par ailleurs, les normes nationales souffrent parfois de leur trop grande généralité. En effet, il est souvent plus pertinent pour les praticiens de prendre des décisions en s'appuyant sur des normes plus spécifiques. Par exemple, pour un psychologue travaillant dans des milieux scolaires socio-économiquement défavorisés, il sera généralement plus utile de disposer de normes élaborées pour ce type de population.

Pour cette dernière raison, mais aussi pour des motifs financiers, il est fréquent de ne développer que des normes locales. Dans ce cas, la population de référence sera plus circonscrite. Elle correspondra, par exemple, aux élèves des écoles de toute une ville ou encore aux patients d'une institution d'accueil pour handicapés. Les normes qui seront générées en référence à ces populations serviront habituellement pour des objectifs très précis : aider à orienter des élèves entre différents établissements, constituer des groupes homogènes pour les apprentissages... Les limites des normes locales découlent de cette très grande spécificité. En effet, pour d'autres usages du test ou du questionnaire, il sera souvent nécessaire de développer de nouvelles normes.

1.2.2 *L'échantillonnage*

Dans le paragraphe précédent, nous avons souligné qu'il n'est généralement pas possible d'établir des normes en testant toute la population de référence. Nous sommes donc contraints d'inférer les caractéristiques de la population à partir des informations contenues dans les résultats d'un échantillon. Les normes ne constituent dès lors qu'une estimation de certains paramètres de la population, comme la moyenne et la variance des scores. Le but de la procédure d'échantillonnage est de minimiser l'erreur d'estimation de ces paramètres. Nous allons passer en revue les principales techniques utilisées pour constituer l'échantillon d'étalonnage.

Pour des raisons d'économie, il est fréquent de recourir à un *échantillon de convenance*. Il est en effet beaucoup plus commode pour le praticien d'utiliser des sujets de son entourage ou des personnes qui se sont présentées volontairement suite à une annonce. Malheureusement, cette procédure d'échantillonnage doit être déconseillée car elle entraîne de sérieux biais dans l'estimation des paramètres de la population. En effet, l'importante place laissée au jugement du praticien ne conduit généralement pas à la constitution d'un échantillon représentatif de la population car les erreurs dues au biais de sélection sont très difficiles à contrôler. De plus, la procédure n'étant pas aléa-

toire, il n'est alors pas possible d'évaluer l'importance de l'erreur d'estimation des paramètres.

Angoff (1971) fait remarquer qu'avec les tests cognitifs, l'usage d'échantillons de convenance conduit habituellement à une surestimation des scores de la population. En effet, les sujets volontaires ou appartenant à l'environnement du chercheur constituent souvent un sous-groupe socio-culturellement favorisé au sein de la population. Mais l'exemple le plus célèbre d'erreur d'estimation due au biais de sélection est certainement celui des sondages précédant l'élection présidentielle américaine de 1948. Tous les instituts de sondage avaient en effet prévu une victoire écrasante de Thomas E. Dewey alors que, finalement, ce fut Harry Truman qui triompha. À cette époque, la technique la plus utilisée était *l'échantillonnage par quotas*. Cette technique consiste à sélectionner l'échantillon de manière systématique afin que ses caractéristiques correspondent exactement à celles de la population. Par exemple, si la population est composée de 49% d'hommes et de 51% de femmes, on demande aux enquêteurs d'interroger des hommes jusqu'au moment où l'échantillon en inclut exactement 49%. La faiblesse de cette méthode d'échantillonnage est de laisser une trop grande place à la subjectivité des interviewers. Un biais, en partie inconscient, risque dès lors d'intervenir dans la sélection des répondants. Ce phénomène s'est d'évidence produit dans les sondages de 1948 et a conduit à une sérieuse mise en question de la méthode d'échantillonnage par quotas.

De manière à contrôler l'erreur d'échantillonnage, c'est-à-dire l'erreur d'estimation des paramètres de la population, il est nécessaire d'exclure toute subjectivité de la procédure et de constituer l'échantillon de manière purement aléatoire. Un échantillon peut être considéré comme aléatoire si chaque sujet de la population a une probabilité égale d'être sélectionné. Si c'est le cas, l'estimation des paramètres de la population sera non biaisée. Par ailleurs, il nous sera possible de calculer l'erreur type d'estimation des différents paramètres et de déterminer un intervalle de confiance autour des valeurs calculées à partir des scores de l'échantillon. Il existe divers techniques d'échantillonnage aléatoire : (1) l'échantillonnage aléatoire simple, (2) l'échantillonnage aléatoire stratifié, (3) l'échantillonnage systématique et (4) l'échantillonnage par grappes.

Nous parlons d'*échantillonnage aléatoire simple* si, d'une population de taille N, nous tirons un échantillon de taille n de telle manière que chaque individu de la population a la même probabilité d'être sélectionné. La procédure d'échantillonnage aléatoire simple consiste à assigner un nombre spécifique à chaque individu de la population puis à tirer au sort parmi les nombres un échantillon dont la taille a été définie au préalable. Pour réaliser ce tirage au sort, nous pouvons utiliser soit une table de nombres aléatoires, soit le générateur de nombres aléatoires inclus dans la plupart des programmes de statistiques actuels (p.e., SAS, SPSS, Systat...). Les tables de nombres aléatoires sont construites pour obtenir une distribution uniforme. Les programmes statistiques permettent, eux, de générer des nombres aléatoires avec différents types de distribution (distribution normale, distribution de t, distribution de χ^2 ...). Dans ce cas, il sera nécessaire de choisir la procédure générant des nombres aléatoires avec une distribution uniforme.

Une fois l'échantillon constitué, nous faisons passer à chacun des individus le test que nous souhaitons étalonner. Sur base des scores de l'échantillon, nous calculons

les statistiques qui nous intéressent et qui seront considérées comme autant d'estimations des paramètres de la population. Les erreurs d'échantillonnage étant inévitables, il importe également d'évaluer l'erreur d'estimation des paramètres. À titre d'illustration, nous prendrons le cas de la moyenne.

Comme nous l'avons vu dans le chapitre 2, du fait des erreurs aléatoires d'échantillonnage, la moyenne que nous calculons à partir des scores de l'échantillon risque d'être sensiblement différente de celle que nous pourrions calculer à partir de tous les scores de la population. Si nous tirions un grand nombre d'échantillons au sein de la population et que nous calculions à chaque fois la moyenne des scores, les différentes moyennes tendraient à se distribuer normalement et leur moyenne serait égale à la moyenne de la population. L'écart type de cette distribution de moyennes est appelé l'erreur type de la moyenne et se note s_M. À partir de l'échantillon que nous avons sélectionné, cette valeur peut-être estimée de manière non biaisée au moyen de la formule suivante, qui est un cas particulier de la formule 2.2 présentée dans le chapitre 2, section 1.2 :

$$s_M = \sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)} \quad (7.1)$$

s^2 est la variance des scores de l'échantillon,

n est la taille de l'échantillon et

N est la taille de la population.

Dans cette formule, la quantité $(N-n)/N$ est appelée la correction pour population finie. Cette correction prend en compte le fait qu'une estimation basée sur un échantillon de 20 sujets tiré d'une population de 60 sujets contient plus d'information à propos de la population qu'un échantillon de 20 sujets tirés d'une population de 10.000 sujets. Cette correction peut être ignorée lorsqu'elle est supérieure ou égale à 0,95, c'est-à-dire lorsque $n \leq (1/20)N$. Dans ce cas, nous retrouvons la formule 2.2 qui s'écrit de manière plus simple :

$$s_M = \sqrt{\frac{s^2}{n}} \quad (7.2)$$

La connaissance de l'erreur type de la moyenne nous permet de construire un intervalle de confiance autour de la moyenne de l'échantillon. Cet intervalle nous oblige à relativiser la valeur obtenue à partir de l'échantillon et à prendre conscience de l'importance de l'erreur d'estimation de la moyenne. Si nous souhaitons avoir 95% de chance que la moyenne de la population se trouve dans l'intervalle de confiance, il nous suffit de multiplier l'erreur type de la moyenne par 1,96 puis, à l'aide de cette valeur, de déterminer la borne inférieure et la borne supérieure de l'intervalle en la soustrayant et en l'additionnant au score moyen de l'échantillon. Par exemple, si la moyenne de l'échantillon est 53,21 et l'erreur type est 3,20, l'intervalle de confiance de 95% sera égal à [53,21 - (3,20 × 1,96) ; 53,21 + (3,20 × 1,96)], c'est-à-dire [46,94 ; 59,48].

La formule 7.2 permet de nous rendre compte aisément que l'erreur type de la moyenne dépend de deux variables : la variance des scores et la taille de l'échantillon.

Plus la taille de l'échantillon est grande et plus la variance des scores est petite, plus faible est l'erreur type de la moyenne ; c'est-à-dire meilleure est l'estimation de la moyenne de la population. Par ailleurs, partant de cette formule, il est possible de déterminer a priori la taille minimum de l'échantillon d'étalonnage nécessaire pour atteindre un niveau d'erreur d'estimation donné. Cette information est économiquement très utile puisqu'elle nous permet d'obtenir la précision d'estimation souhaitée au moindre coût. La taille de l'échantillon peut être déterminée à l'aide de la formule suivante :

$$n = \frac{N\sigma^2}{ND + \sigma^2} \tag{7.3}$$

N = taille de la population

σ^2 = la variance des scores de la population

$D = \frac{B^2}{4}$

B est la borne de l'erreur d'estimation que nous avons choisie et correspond à deux fois l'erreur type d'estimation. Cette valeur, définie a priori, doit nous permettre de construire un intervalle de confiance de 95% autour de la moyenne de l'échantillon. Quant à la variance des scores de la population, elle nous est inconnue. Il est donc nécessaire d'estimer celle-ci à partir des résultats d'un échantillon. Souvent, les résultats recueillis lors d'une première expérimentation du test sont utilisés à cet effet. La taille de l'échantillon devra cependant être suffisante pour permettre une estimation assez précise de la variance de la population.

Nous pouvons illustrer l'utilisation de la formule 7.3 par l'exemple d'un test d'orthographe, constitué de 80 mots d'usage, que nous souhaitons étalonner pour la classe de 4e année primaire française. En 1993-94, la population de 4e année primaire était de 117.395 élèves. Nous désirons déterminer la taille minimale de l'échantillon nécessaire pour estimer la moyenne des scores de cette population avec une marge d'erreur égale à 2 points (en plus ou en moins). Une première expérimentation du test sur un échantillon de 75 élèves a permis d'estimer la variance des scores de la population, laquelle est approximativement égale à 225. Par conséquent :

$$B = 2$$

$$D = \frac{2^2}{4} = 1$$

$$n = \frac{117395 \times 225}{(117395) \times 1 + 225} = 225$$

Il faudrait donc sélectionner un échantillon aléatoire simple de 225 élèves de 4ème année primaire pour estimer la moyenne de la population au test d'orthographe avec 95% de chance que la moyenne de la population soit incluse dans l'intervalle de ± 2 points autour de la moyenne de l'échantillon. Si nous désirons que cet intervalle soit de ± 1 point, la taille de l'échantillon devra être au minimum de 893 élèves. Nous constatons que, dans ce cas, l'amélioration de la précision implique une augmentation très importante de la taille de l'échantillon nécessaire.

L'échantillonnage aléatoire stratifié consiste à rassembler les individus de la population au sein de groupes sans recouvrement, appelés strates, et à ensuite sélectionner un échantillon aléatoire simple dans chacune des strates ainsi constituées. Par exemple, pour étalonner un test de mémoire de séries de chiffres, nous diviserons la population en cinq groupes définis par le niveau d'étude puis nous tirerons au hasard dans chaque groupe un nombre d'individus proportionnel à l'importance de ce groupe au sein de la population. Dans ce cas, aucune des strates ne se recouvrent puisqu'un individu ne peut appartenir qu'à une et une seule strate. Par conséquent, les échantillons sélectionnés dans les différentes strates seront indépendants les uns des autres.

Le principal avantage de l'échantillonnage aléatoire stratifié est de permettre une estimation des paramètres de la population plus précise que celle obtenue avec un échantillon aléatoire simple de même taille. Cet avantage n'est cependant effectif que si la population est divisée en strates relativement homogènes sur base d'une ou plusieurs variables corrélées avec la variable mesurée par le test. C'est le cas dans notre exemple puisque la mémoire de séries de chiffres est corrélée avec le niveau scolaire. La variance au sein de chaque strate est dès lors plus faible que la variance au sein de la population. Dans cet exemple, il est possible d'encore réduire la variance intra-strate en définissant chacune de celles-ci sur base des variables “ niveau scolaire ” et “ âge ”. En effet, l'âge est également corrélé avec la mémoire de séries de chiffres. Les strates définies en tenant compte du niveau scolaire et de l'âge seront dès lors plus homogènes que celles définies en tenant compte de l'âge seul.

Un second avantage de l'échantillonnage aléatoire stratifié est de nous permettre d'estimer aisément les paramètres de sous-groupes de la population. Nous pouvons par exemple estimer le score de mémoire moyen selon l'âge et selon le niveau scolaire. Enfin, un dernier avantage de l'échantillonnage aléatoire stratifié est de donner plus de crédibilité aux normes. Les utilisateurs de tests accordent en effet une plus grande confiance à des normes basées sur un échantillon qui respecte la composition démographique de la population, même si certaines caractéristiques de cette dernière ne sont nullement corrélées avec la variable mesurée par le test.

Avec un échantillon aléatoire stratifié, le calcul de l'erreur d'estimation est plus complexe qu'avec un échantillon aléatoire simple. L'erreur type de la moyenne peut être estimée à l'aide de la formule suivante :

$$\hat{s}_M = \sqrt{\frac{1}{N^2}\sum N_i^2\left(\frac{N_i - n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)} \tag{7.4}$$

N = la taille de la population

N_i = la taille de la *i-ème* strate au sein de la population

n_i = la taille de l'échantillon tiré au sein de la *i-ème* strate

s_i^2 = la variance des scores de l'échantillon tiré au sein de la *i-ème* strate

Le nombre d'individus dans une strate de la population affecte la quantité d'information incluse dans un échantillon tiré au sein de cette strate. Par conséquent, la taille de l'échantillon tiré dans chaque strate est habituellement proportionnelle à la taille de la strate au sein de la population. Si ce principe est respecté, l'estimation de la

moyenne à partir d'un échantillon aléatoire stratifié équivaut à celle estimée à partir d'un échantillon aléatoire simple.

L'échantillonnage systématique consiste à choisir de manière aléatoire un individu au sein d'une liste (ou de tout autre cadre de référence) puis, à partir de celui-ci, de sélectionner tous les *k-ièmes* individus de la liste. Par exemple, un praticien qui souhaite étalonner un test d'intelligence au sein d'une école utilisera la liste alphabétique des élèves au sein de laquelle il choisira de manière aléatoire un premier sujet. À partir de celui-ci, il sélectionnera systématiquement tous les 10ièmes sujets tout au long de la liste jusqu'à la fin de celle-ci.

Le principal avantage de l'échantillonnage systématique réside dans sa facilité de mise en oeuvre. L'échantillonnage aléatoire simple et l'échantillonnage aléatoire stratifié représentent des procédures nettement plus coûteuses en temps. Il faut en effet numéroter tous les individus de la population avant de réaliser un tirage aléatoire. Cette procédure est particulièrement laborieuse lorsque la taille de la population est très grande et elle est même impossible lorsque nous ne connaissons pas la taille de la population et/ou que nous n'en possédons pas de liste exhaustive. Dans ce cas, l'échantillonnage systématique se révèle une procédure de choix. En effet, nous pouvons sélectionner les sujets à partir d'une liste (par exemple un fichier alphabétique ou un annuaire téléphonique) mais nous pouvons aussi les choisir en possédant seulement une définition en compréhension, mais non en extension, de la population de référence. Par exemple, un psychologue peut étalonner un questionnaire de dépression, destiné aux patients de l'hôpital où il travaille, en le faisant passer par un individu sur trois vus en consultation. Dans ce cas, la taille de la population est inconnue et aucune liste des individus n'est évidemment disponible. Lorsque nous possédons une liste exhaustive de la population, la règle pour déterminer la périodicité de la sélection est de choisir une valeur k plus petite ou égale au rapport entre la taille de la population et la taille de l'échantillon (c'est-à-dire $k \leq N/n$). Par exemple, si la population est égale à 2000 et l'échantillon est égal à 50, k devra être égal ou inférieur à 40.

Pour un échantillon systématique, l'estimation de l'erreur type de la moyenne se calcule selon la même formule que pour l'échantillon aléatoire simple (formule 7.1). Lorsque la taille de la population est inconnue, la formule 7.2 devra être utilisée. Toutefois, l'identité des formules utilisées n'implique pas que l'estimation de la moyenne de la population est similaire avec les deux procédures d'échantillonnage. En réalité, elle n'est équivalente que si la liste des individus de la population est aléatoire, c'est-à-dire si la corrélation est nulle entre le critère d'organisation de la liste et la variable mesurée par le test. Par exemple, la corrélation entre le classement alphabétique des élèves et leur niveau d'intelligence est certainement égale à zéro, ce qui nous permet de considérer le fichier alphabétique des élèves comme une liste aléatoire. Le praticien devra être attentif à cette question car la succession des individus n'est pas toujours indépendante de la variable mesurée et l'estimation des paramètres de la population risque alors d'être biaisée. Par exemple, le niveau moyen de dépression peut varier en fonction des périodes de l'année. Par conséquent, le psychologue qui sélectionne de manière systématique un échantillon sur une période restreinte risque d'obtenir des normes biaisées.

L'échantillonnage en grappes consiste en la sélection aléatoire de collections de sujets, appelées *grappes*. L'unité tirée au sort n'est dès lors plus un individu mais un

ensemble d'individus. Cette technique est, dans un certain nombre de situations, la moins coûteuse à mettre en oeuvre. C'est le cas lorsque nous ne possédons pas de liste des individus de la population et/ou que le testing individuel de chaque sujet se révèle difficile. Imaginons, par exemple, un test d'acquis scolaires devant être étalonné pour des élèves de l'enseignement secondaire. Si nous utilisons une des techniques d'échantillonnage précédemment décrites, nous allons devoir extraire un ou deux élèves d'un grand nombre de classes dans le but de les tester. Cette procédure est évidement laborieuse et perturbante pour le bon fonctionnement des classes. Dans ce cas, il est souvent plus simple de sélectionner aléatoirement des classes entières et de tester tous les élèves qui en font partie.

L'échantillonnage en grappes est d'autant plus efficace que les grappes sont hétérogènes. Nous récoltons alors un maximum d'information à propos de la population au moindre coût. Par contre, lorsque les grappes sont très homogènes, nous sommes obligés de tester un grand nombre d'individus pour recueillir une information relativement limitée. Par ailleurs, pour que notre échantillon soit aléatoire, il est nécessaire de pouvoir constituer une liste de toutes les grappes de la population. Nous pourrons alors tirer au sort un échantillon de grappes en utilisant la technique décrite pour l'échantillonnage aléatoire simple. Un échantillon par grappes ne nous donne une estimation non biaisée de la moyenne de la population qu'à condition que les grappes soient de tailles identiques et qu'elles soient suffisamment nombreuses. Ces conditions sont souvent difficiles à remplir, ce qui risque d'entraîner des biais d'estimation des normes. Pour un échantillon par grappes, l'estimation de l'erreur type de la moyenne peut être calculée au moyen de la formule suivante :

$$\hat{S}_M = \sqrt{\left(\frac{N-n}{Nn\overline{M}^2}\right)\frac{\sum (y_i - \bar{y}m_i)^2}{n-1}} \qquad (7.5)$$

N = le nombre de grappes dans la population

n = le nombre de grappes sélectionnées dans l'échantillon

$\overline{M}$ = la taille moyenne des grappes dans la population

y_i = le total des scores dans la *i-ème* grappe

m_i = le nombre de sujets dans la *i-ème* grappe

$\bar{y}$ = la moyenne des scores de l'échantillon

Une question que se posent fréquemment les praticiens lorsqu'ils définissent les normes d'un test concerne le niveau acceptable des erreurs d'estimation. De la réponse à cette question découle la détermination de la taille de l'échantillon nécessaire pour établir les normes. Comme le souligne justement Angoff (1971, p.558), *" on ne peut malheureusement pas répondre à cette question dans l'abstrait "*. Le niveau acceptable des erreurs d'estimation dépend en effet de l'usage qui sera fait des normes et du coût que nous sommes prêts à consacrer à la constitution de ces dernières. Or l'arbitrage entre la précision et le coût est toujours un problème spécifique à chaque situation. Le praticien doit en priorité prendre en compte l'importance des décisions qui seront prises sur base des normes et mettre en balance le coût d'une mauvaise décision et le coût d'une augmentation de la précision des normes. Pour aider le praticien,

Angoff (1971) propose de prendre également en compte une règle simple : l'erreur type de la moyenne ne devrait pas excéder de 14% l'erreur type de mesure des résultats à un test.

1.3 La transformation des scores

Les résultats recueillis sur l'échantillon d'étalonnage ne sont habituellement pas communiqués tels quels. Pour permettre une interprétation plus aisée des résultats de tests, les scores bruts de l'échantillon d'étalonnage sont généralement transformés et présentés sur une échelle familière aux praticiens. Il existe de nombreuses échelles destinées à exprimer les normes. Nous ne présenterons ici que les plus courantes. À chaque fois, nous expliciterons la procédure de transformation des scores puis nous discuterons des avantages et inconvénients de l'échelle en question.

1.3.1 *Les échelles en niveaux d'âge*

Les normes peuvent être exprimées en termes d'âges moyens auxquels diverses performances sont réussies. Les sujets testés se verront alors attribuer un niveau d'âge en fonction de leurs résultats bruts. L'étalonnage en niveaux d'âge se déroule généralement selon les étapes suivantes :

1. Des échantillons de sujets pour les âges considérés sont constitués. Habituellement, un âge est défini comme un intervalle plus ou mois large autour de l'âge en question. Par exemple, un échantillon d'enfants de six ans comprendra des sujets âgés de 6 ans plus ou moins 2 mois, c'est-à-dire des sujets situés dans l'intervalle 5 ans 10 mois - 6 ans 2 mois.
2. Le score moyen de chaque groupe d'âge est calculé.

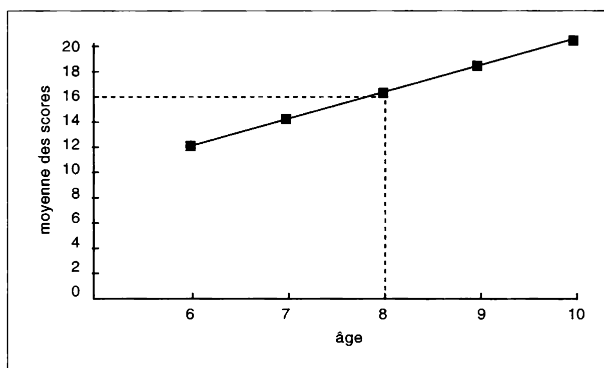


Figure 1 – Interpolation d'un niveau d'âge

3. Eventuellement, les scores de certaines tranches d'âge sont estimés par interpolation (figure 1). Cette procédure est utilisée lorsque certains âges n'ont pas été inclus dans l'échantillon d'étalonnage ou que l'on désire présenter des normes par mois, et non seulement par année. La procédure d'interpolation repose sur le postulat d'une progression linéaire des caractéristiques évaluées par le test. Elle consiste à calculer la valeur intermédiaire entre les valeurs recueillies sur

l'échantillon d'étalonnage. Par exemple, si le score moyen obtenu au test par les enfants de 7 ans est de 14 points et que celui obtenu par les enfants de 9 ans est de 18 points, on peut estimer que le score moyen des enfants de 8 ans est de 16 points.

Le niveau d'âge le plus connu est certainement l'âge mental qui représente le niveau de développement intellectuel atteint par un sujet. Mais on peut utiliser le principe du niveau d'âge pour caractériser n'importe quelle capacité ou aptitude (la motricité, la connaissance du schéma corporel...) pour peu que celle-ci varie avec l'âge. Ceci représente d'ailleurs la limite essentielle de l'expression des normes en niveaux d'âge. Ce principe est en effet inapplicable lorsque le trait mesuré ne varie pas spécifiquement avec l'âge (par exemple l'anxiété) ou lorsqu'il est arrivé au terme de son développement (par exemple l'intelligence adulte). Par ailleurs, même lorsque le trait mesuré évolue avec l'âge, la corrélation entre les variables " âge " et " performance " est rarement parfaite. Pour que ce soit le cas, il faudrait que la relation entre l'évolution de l'âge et celle de la performance soit rigoureusement linéaire, ce qui n'est pas toujours le cas. Durant l'enfance, les progrès ne sont en effet pas proportionnels à l'âge et se font à des rythmes variés. Le lien entre âge et performance est donc assez lâche. Par conséquent, le niveau d'âge attribué à une performance sera plus ou moins adéquat en fonction du degré de corrélation linéaire entre ces deux variables.

Un autre problème soulevé par les niveaux d'âge concerne leur interprétation. Ainsi, les praticiens ont souvent tendance à assimiler le raisonnement de tous les sujets de même âge mental. En fait, cette assimilation n'est pas en accord avec la réalité psychologique. Un adulte handicapé dont l'âge mental est de 8 ans ne réfléchit pas comme un enfant de 8 ans de même âge mental. Dans le premier cas, nous avons affaire à une pensée figée, marquée par les stéréotypies, alors que, dans le second cas, il s'agit d'une intelligence mobile dont l'évolution n'est pas achevée. Les performances des deux sujets sont quantitativement semblables mais les compétences sous-jacentes sont loin d'être identiques du point de vue qualitatif.

Enfin, un dernier problème posé par l'utilisation des niveaux d'âge provient de la relativité des unités d'âge. Ainsi, un retard d'un an à 4 ans (âge chronologique) n'a pas la même valeur qu'un retard d'un an à 12 ans. Le même problème se pose de façon plus illustrative au niveau de la taille. Une différence de 5 cm est en effet beaucoup plus importante entre deux nouveau-nés qu'entre deux adultes.

Pour résoudre cette difficulté, nous pouvons calculer un quotient en divisant le niveau d'âge par l'âge réel du sujet. Il est possible de calculer non seulement des quotients intellectuels mais aussi des quotients de développement moteur, des quotients de mémoire... De cette façon, nous évitons de considérer le niveau d'âge comme une valeur absolue et nous l'interprétons comme une valeur relative à l'âge chronologique. Si le niveau d'âge d'un sujet évolue parallèlement à son âge chronologique, alors son quotient restera constant au cours du développement. Toutefois, cette façon de procéder ne doit pas nous faire oublier que le rapport ainsi calculé s'appuie sur une mesure en niveaux d'âge dont nous avons souligné les sérieuses faiblesses.

1.3.2 *Les échelles en niveaux scolaires*

L'expression des normes en niveaux scolaires a d'importantes similitudes avec celle en niveaux d'âge. La procédure d'étalonnage est en effet semblable, à cette différence près que nous constituons des groupes de niveau scolaire (par exemple, 1ère primaire, 2e primaire...) au lieu de groupes d'âge. Une performance sera dès lors caractérisée par l'année scolaire où elle est atteinte par la moyenne des élèves. Nous considérerons, par exemple, qu'un score brut donné est du niveau de la 4e année primaire s'il est obtenu par la moyenne des élèves de cette année scolaire.

Les désavantages des niveaux scolaires sont similaires à ceux des niveaux d'âge. Derrière la simplicité apparente de l'interprétation se cachent en effet les mêmes problèmes, mais accentués. Parmi ceux-ci, le plus fondamental est que la corrélation entre les niveaux scolaires et les niveaux de performance est loin d'être parfaite. Pour que ce soit le cas, il faudrait admettre que l'évolution des acquis est régulière et continue tout au long de l'année, ce qui est peu vraisemblable. Il faudrait également que la variabilité des performances entre les classes et les établissements scolaires soit faible. Or, c'est le phénomène inverse qui est régulièrement observé : le niveau moyen de performance varie fortement d'une école à l'autre, et même d'une classe à l'autre. Cette variabilité est due aux caractéristiques sociologiques des populations de chaque école mais aussi aux différences de pratiques d'enseignement et de promotion (redoublements fréquents ou non) entre les écoles. En découlent des recouvrements importants entre les performances des élèves des différentes années. Dans ces conditions, prendre comme référence le niveau moyen de performance correspondant à un niveau scolaire précis conduit souvent à d'importantes erreurs d'appréciation et à des prises de décision inadéquates.

1.3.3 *Les échelles en rangs centiles*

La valeur d'un résultat peut être exprimée en terme de place ou de rang au sein de la population. Les centiles (ou percentiles) sont une des modalités les plus fréquentes de graduation des rangs. La distribution des résultats bruts est alors ramenée à 100 échelons afin qu'entre chaque échelon se trouve 1% des sujets. Cette transformation des résultats en centiles s'appelle le centilage. La procédure de calcul des rangs centiles est présentée en détail dans le §2 du premier chapitre.

Chaque valeur de la distribution est prise comme ordinale et non comme cardinale. Par exemple, le centile 80 indique la 80e place et non 80 points. Dans ce cas, 80% des sujets ont des résultats bruts inférieurs à celui de l'individu testé. Plus faible sera le résultat d'un sujet, plus bas sera le centile et inversement. N'oublions donc pas que, contrairement aux places d'examens scolaires (la première place est attribuée au meilleur résultat), dans une échelle en centiles, le premier rang est donné au score brut le plus faible et inversement.

Dans la pratique, il n'est pas toujours nécessaire ni possible d'établir 100 divisions, soit que la variable a moins d'extension, soit qu'une discrimination aussi détaillée n'est pas nécessaire. On peut alors utiliser une notation en déciles (10 rangs) ou en quartiles (4 rangs).

L'expression des normes en centiles (ou en déciles ou en quartiles) présente un important inconvénient. Une telle distribution des notes est rectangulaire alors que la distribution des notes brutes est généralement normale. Autrement dit, la transformation en centiles ne respecte pas la forme de la distribution originelle et modifie donc les rapports entre les résultats. Le problème apparaît clairement sur la figure 2. Au voisinage de la moyenne, les sujets sont nombreux et, par conséquent, les centiles sont très proches. Par contre, aux extrémités de la distribution, les sujets se raréfient et les centiles sont donc de plus en plus éloignés les uns des autres. Ainsi, l'écart entre le centile 50 et le centile 60 n'est pas égal à l'écart entre le centile 80 et le centile 90. Il en découle un sérieux problème de comparaison entre sujets. Les centiles ne nous renseignent donc que sur le rang d'un sujet mais non sur l'écart qui le sépare des autres sujets. N'oublions pas que nous avons affaire à une échelle ordinale avec toutes les limites statistiques que cela représente.

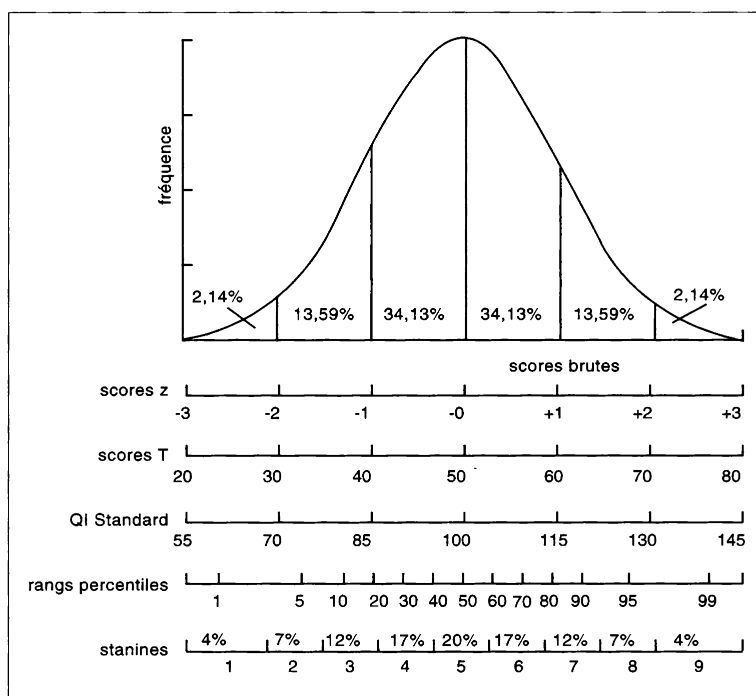


Figure 2 – Relation entre la courbe normale et les différents types de scores transformés

1.3.4 *Les échelles en scores standard*

La transformation en scores standard résout différents problèmes rencontrés avec les centiles. Cette transformation ne modifie pas la forme de la distribution des scores bruts car elle préserve au sein de la nouvelle distribution les relations numériques existant dans la distribution originelle (figure 2). En effet, pour chaque valeur de la distribution des notes brutes, nous ne faisons que retrancher une constante ($\bar{X}$) et diviser par une constante (s). Si, pour une distribution donnée, nous représentons graphiquement les coordonnées entre chaque score brut et les scores standard correspon-

dant, nous pourrons tracer une ligne droite passant exactement par tous les points ainsi représentés. Pour cette raison, la transformation en scores standard est qualifiée de linéaire puisqu'elle est du type $Y = aX + b$ (voir chapitre 1, §1). Les échelles en scores standard peuvent être considérées comme des échelles d'intervalle et possèdent donc les avantages de ces dernières concernant les traitements statistiques. Comment s'effectue pratiquement la transformation des scores bruts en scores standard ?

Il est tout d'abord nécessaire de calculer la moyenne et l'écart type de la distribution des notes brutes. Connaissant ces valeurs, nous pouvons alors transformer chaque score brut en calculant la distance qui le sépare de la moyenne avec une unité égale à l'écart type. Nous obtenons ainsi des scores z, dont nous avons déjà parlé dans le §3 du premier chapitre. La formule de transformation en score z est la suivante :

$$z_i = \frac{X_i - \bar{X}}{s} \tag{7.6}$$

Par exemple, si dans une distribution de scores bruts $\bar{X} = 60$ et $s = 5$ alors :

pour $X = 65$ $\quad z = \frac{65 - 60}{5} = +1{,}00$

pour $X = 58$ $\quad z = \frac{58 - 60}{5} = -0{,}40$

Les scores z ont comme inconvénient de présenter des décimales et d'être de signe négatif pour tous les scores inférieurs à la moyenne. C'est pourquoi il est d'usage d'utiliser une moyenne et un écart type arbitraires qui permettent de transformer les scores bruts en des valeurs entières et positives. Concrètement, la procédure consiste à multiplier chaque score z par un même écart type puis à lui ajouter une même valeur moyenne. Soulignons que cette procédure préserve le caractère linéaire de la transformation. La formule 7.6. devient alors :

$$Y_i = s'z_i + \bar{X}' \tag{7.7}$$

Ce qui peut s'exprimer de manière plus détaillée :

$$Y_i = s'\left(\frac{X_i - \bar{X}}{s}\right) + \bar{X}' \tag{7.8}$$

Dans ces deux formules, s' et $\bar{X}'$ sont respectivement les valeurs arbitraires de l'écart type et de la moyenne. Il existe quelques valeurs courantes pour s' et $\bar{X}'$:

- pour la transformation en *score T*, $s' = 10$ et $\bar{X}' = 50$
- pour la transformation en *Q.I. standard*, utilisée dans les tests de Wechsler, $s' = 15$ et $\bar{X}' = 100$
- pour la transformation en *score CEEB* (College Entrance Examination Board), $s' = 100$ et $\bar{X}' = 500$

À titre d'exemple, reprenons les données présentées ci-dessus et transformons-les en *scores T* :

$$\text{pour } X = 65 \quad Y = 10\left(\frac{65-60}{5}\right) + 50 = 60$$

$$\text{pour } X = 58 \quad Y = 10\left(\frac{58-60}{5}\right) + 50 = 46$$

Comme nous venons de le voir, la transformation en score standard présente des avantages certains. Elle demande toutefois aux praticiens d'être attentifs à la valeur de écart type utilisée pour la transformation. De grossières erreurs d'interprétation des scores peuvent en effet découler d'une méconnaissance de cette valeur. Par exemple, nous avons vu que les tests de Wechsler utilisent une moyenne de 100 et un écart type de 15. Par contre, Cattell, pour le *Culture Free Test*, utilise une moyenne de 100 et un écart type de 24. Par conséquent, un sujet qui se situe à un écart type en dessous de la moyenne aura 85 de QI au test de Wechsler et 76 de QI au *Culture Free Test.* La différence entre les scores transformés est importante alors que la position du sujet dans la distribution des scores bruts est identique. On conçoit aisément le type d'erreur qui pourrait être commise par simple ignorance des caractéristiques de l'échelle sur laquelle sont présentées les normes.

1.3.5 *Les échelles en scores standard normalisés*

Nous avons vu que la transformation en score standard est une transformation linéaire qui ne modifie pas la forme de la distribution des scores bruts. Cependant, il est parfois raisonnable de penser que le trait mesuré se distribue normalement et que la non normalité de la distribution des scores bruts résulte d'erreurs d'échantillonnage. Ainsi, les constructeurs de tests d'intelligence s'appuient généralement sur le postulat d'une distribution normale de l'intelligence au sein de la population. Dans ce cas, il est d'usage d'effectuer une transformation en scores standard qui normalise la distribution des scores bruts. Cette transformation est intéressante car la distribution normale possède des caractéristiques bien connues et les résultats sont dès lors plus faciles à interpréter. Nous avons vu dans le §3 du premier chapitre que, pour chaque valeur de cette distribution, nous connaissons précisément le pourcentage de cas qui se situent au-dessous et au-dessus. Dès lors, les comparaisons entre les résultats de différents tests sont grandement facilitées, à condition que la transformation en scores normalisés ait été réalisée, dans chaque cas, en utilisant une même moyenne et un même écart type. Nous savons alors qu'un sujet qui a obtenu un même score standard à deux tests différents occupe exactement la même position au sein de la distribution des scores de ces tests.

Puisqu'elle modifie la forme de la distribution d'origine, la transformation en score standard normalisé est non linéaire. La technique de transformation la plus simple se fait en deux étapes. Les scores bruts sont tout d'abord transformés en percentiles en utilisant la formule présentée dans le §2 du premier chapitre. Les centiles ainsi obtenus sont ensuite transformés en scores z à l'aide de la table de la distribution normale réduite. Par exemple, si un score brut correspond au centile 80, nous chercherons dans la table de la distribution normale réduite la valeur de z sous laquelle se trouvent 80% des cas. En l'occurrence, cette valeur est égale à 0,84. Si nous opérons de la sorte pour tous les scores bruts de la distribution, nous ferons correspondre à chacun de ceux-ci des scores z dont la distribution sera parfaitement normale. Pour éviter les valeurs décimales et négatives des scores z, il nous suffira d'appliquer la formule de transfor-

mation présentée plus haut en utilisant une moyenne et un écart type adéquat. Dans notre exemple, nous pourrions ainsi faire correspondre à 0,84 la valeur 113 au sein d'une distribution dont la moyenne est égale à 100 et l'écart type est égal à 15.

Dans certains cas, la transformation que nous venons de détailler procure une échelle inutilement fine pour l'usage auquel le test est destiné. Par exemple, si nous souhaitons évaluer le niveau de connaissance que possèdent des adultes en anglais afin de les orienter vers différents programmes de perfectionnement, nous n'aurons pas besoin d'un test gradué en cent échelons. La normalisation se fait alors selon un nombre de catégories plus limité. La transformation en stanine en est un exemple bien connu (figure 2). Dans ce cas, les scores standard normalisés sont limités à 9 avec une moyenne égale à 5 et un écart type approximativement égal à 2. Le *stanine (standard nine)* au centre la distribution (des rangs centiles 40 à 59) contient 20% des cas. Le premier et le dernier stanine contiennent 4% des cas, le second et le huitième 7%, le troisième et le septième 12%, le quatrième et le sixième 17%.

La figure 3 illustre la relation existant entre les scores bruts et les rangs centiles. Lorsque la distribution des scores bruts est parfaitement normale, cette relation prend la forme d'une ogive normale. Mais habituellement, du fait d'erreurs d'échantillonnage, les coordonnées entre les scores bruts et les rangs centiles ne correspondent qu'approximativement à cette courbe. Cela signifie qu'en fonction des échantillons, un même score brut peut correspondre à différents rangs centiles. Pour atténuer cet effet de l'erreur d'échantillonnage, nous pouvons recourir à la *technique du lissage* qui complexifie quelque peu la procédure de normalisation. La manière la plus rigoureuse d'effectuer le lissage constitue à déterminer la fonction mathématique qui s'ajuste le mieux aux coordonnées entre scores bruts et rangs percentiles. Sur base de cette fonction, nous pouvons alors tracer la courbe lissée. Les différents points de cette courbe constituent les nouvelles coordonnées entre les scores bruts et les rangs centiles. Nous obtenons ainsi une correspondance entre les deux ensembles de valeurs purifiées des erreurs d'échantillonnage. À titre d'exemple, nous avons représenté sur la figure 3 la relation entre un score brut de 40 et le centile 50 qui lui correspond. Partant des rangs centiles, nous déterminons ensuite les scores z (et éventuellement des scores standard) selon la procédure décrite plus haut.

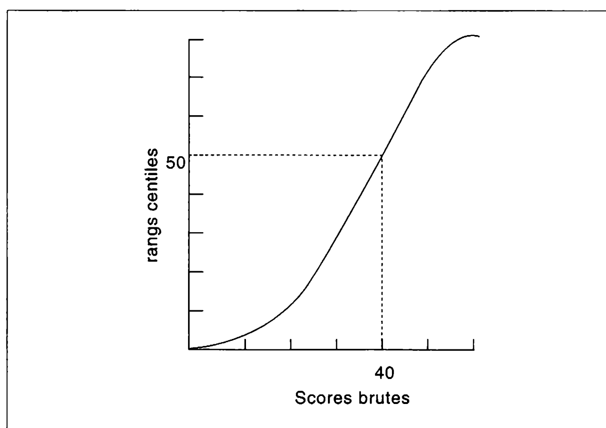


Figure 3 – Représentation graphique de la relation entre scores bruts et rangs centiles

La procédure de lissage suppose que la distribution des scores bruts de l'échantillon ne s'écarte pas trop de celle de la population. En effet, si l'écart est important, la courbe lissée risque d'être mal estimée et de s'écarter sensiblement de celle qui pourrait être tracée sur base des scores de la population. Par conséquent, l'utilisation de la procédure de lissage suppose un échantillonnage rigoureux en vue de minimiser les erreurs d'estimation.

D'une manière générale, la transformation en scores standard normalisés n'est envisageable que si la distribution des scores bruts est relativement proche de la distribution normale. Par ailleurs, l'existence au sein de la population d'une distribution normale du trait mesuré doit être conceptuellement plausible. Par exemple, il est logique que la distribution des scores à un questionnaire de dépression présente une forte asymétrie négative. La majorité des sujets tout-venant auront en effet un score de dépression très faible. Par contre, les sujets déprimés auront des scores qui s'étaleront vers la droite de la distribution. Normaliser une telle distribution n'aurait aucun sens. La procédure de normalisation des scores ne peut donc pas être appliquée de manière automatique. Elle doit s'appuyer sur une analyse détaillée de la distribution des scores bruts et sur une compréhension effective de la réalité mesurée.

2. Équivalence entre les scores de différents tests

2.1 Conditions pour la mise en équivalence de scores

Le praticien peut être amené à comparer des résultats obtenus à différents tests mesurant une même réalité. Pour effectuer une telle comparaison, il est nécessaire de rendre équivalents les scores aux tests considérés. Ceci implique que les scores d'un des tests soient convertis dans la métrique de l'autre test. On note habituellement y* les scores au test X convertis dans la métrique du test Y. La mise en équivalence des scores de deux tests est qualifiée d'*horizontale* lorsque ces tests ont le même degré de difficulté. On parlera de mise en équivalence *verticale* lorsque les deux tests ont des niveaux de difficulté différents. C'est le cas lorsque l'on veut mettre en équivalence les résultats de tests d'aptitude construits pour évaluer des sujets appartenant à différentes tranches d'âges.

Le principe général de la mise en équivalence peut être illustré par la conversion des degrés Fahrenheit en degrés Celsius. Dans ce cas, les deux thermomètres utilisés mesurent une même réalité, à savoir la température, mais sur des échelles différentes. La conversion en degrés Celsius (°C) des températures relevées en degrés Fahrenheit (°F) s'effectue en retirant " 32 " de la température exprimée en Fahrenheit et en multipliant le résultat par 5/9. Selon cette formule de conversion, 50° F sont ainsi équivalents à 10° C. Une fois la conversion effectuée, toutes les températures enregistrées initialement en degrés Fahrenheit sont strictement équivalentes à celles enregistrées en degrés Celsius. Du fait de cette possibilité de mise en équivalence, il est indifférent d'observer les températures à l'aide d'un thermomètre gradué en degrés Fahrenheit ou en degrés Celsius.

Cette propriété des scores mis en équivalence se retrouve dans le cas des tests. En effet, selon Lord (1977, p.128)," *des scores transformés y* et des scores bruts x*

peuvent être qualifiés d'équivalents si et seulement s'il est indifférent que les sujets soient évalués avec le test X ou le test Y ". Pour que les scores à deux tests X et Y puissent être mis en équivalence, un certain nombre de conditions doivent dès lors être remplies (Lord, 1980) :

(1) Les deux tests doivent mesurer la même caractéristique.

(2) Les mesures réalisées avec les deux tests doivent être équitables. Cela signifie que certains sujets ne doivent pas être défavorisés en passant le test X plutôt que le test Y, et réciproquement. Pour que cette équité soit garantie, il est nécessaire que le score vrai d'un sujet au test Y soit identique à son score vrai au test Y* (c'est-à-dire au test X dont les scores ont été convertis dans la métrique du test Y). Il faut également que l'erreur de mesure soit égale au test Y et au test Y*.

(3) La conversion doit être indifférente aux groupes qui ont servi à élaborer les tables de transformation des scores.

(4) La conversion doit être symétrique. Cela signifie qu'il est indifférent de réaliser la transformation du test X vers le test Y ou du test Y vers le test X.

Ces conditions sont d'évidence difficiles à satisfaire dans la pratique. Pour que ce soit le cas, il faudrait que les tests X et Y soient strictement parallèles, ce qui est pratiquement impossible. En particulier, la seconde des quatre conditions est sans doute celle qui soulève le plus de problèmes (Petersen & al., 1989). Il est en effet peu vraisemblable de pouvoir construire deux tests dont la fiabilité serait égale à tous les niveaux d'aptitude et qui présenteraient dès lors des distributions de fréquences conditionnelles identiques. Certains psychométriciens (Morris, 1982) ont donc suggéré de remplacer cette condition d'équité forte par une condition d'équité faible. Selon celle-ci, seul le score moyen conditionnel doit être identique au test Y et au test Y*. En d'autres termes, le score attendu d'un sujet doit être le même avec le test Y qu'avec le test Y*. Cette exigence, certainement plus réaliste, rend possible des mises en équivalence dans le cadre de la théorie classique des scores.

Toutefois, nous devons reconnaître que, dans les faits, ces mises en équivalence restent souvent approximatives car les conditions requises ne sont qu'imparfaitement remplies. Nous sommes ici confrontés aux limites de la théorie classique. Nous verrons dans le chapitre 8 que les modèles de la réponse à l'item apportent des solutions certainement plus satisfaisantes aux problèmes de mise en équivalence. Nous présentons cependant les deux techniques de mise en équivalence les plus fréquemment utilisées dans le cadre de la théorie classique car elles sont les seules applicables lorsqu'on ne dispose que de petits échantillons. Ces deux techniques sont la mise en équivalence linéaire et la mise en équivalence équipercentile.

2.2 La mise en équivalence linéaire

Cette technique est basée sur le postulat d'une relation linéaire entre les scores au test X et au test Y. On suppose alors que les distributions des scores aux deux tests ne diffèrent que par leurs moyennes et leurs écarts type. Si c'est le cas, nous pouvons écrire que :

$$y = ax + b \tag{7.9}$$

où
$$a = \frac{S_Y}{S_X}$$

$$b = \bar{Y} - \frac{S_Y}{S_X}\bar{X}$$

La formule de conversion des scores au test X dans la métrique du test Y s'écrit dès lors :

$$y^* = \frac{S_Y}{S_X}(x - \bar{X}) + \bar{Y} \tag{7.10}$$

Puisque la relation entre les deux distributions est linéaire, des scores équivalents au test X et au test Y correspondent au même score z. Il est par conséquent possible de réaliser la conversion des scores entre les deux tests via la correspondance des scores z.

La mise en équivalence linéaire peut se faire en utilisant divers plans expérimentaux. Le plus simple consiste à faire passer les deux instruments à un même groupe de sujets. Souvent, l'ordre de passage de chaque test est déterminé de manière aléatoire pour éviter un effet d'ordre. L'inconvénient de cette procédure est d'être assez lourde pour les sujets qui, tous, doivent passer les deux tests. Pour cette raison, on préfère parfois utiliser un autre plan expérimental où chacun des tests est passé par un groupe différent de sujets. Pour ces derniers, la procédure est ainsi plus légère. Pour les praticiens par contre, elle exige de constituer les groupes de manière strictement aléatoire afin de garantir leur équivalence statistique.

Un troisième plan expérimental consiste à faire passer chaque test à des groupes différents tout en administrant à chacun de ceux-ci une épreuve commune relativement courte, qualifiée de test d'ancrage. L'intérêt de cette procédure est de maintenir dans des limites raisonnables le temps de passation de chaque sujet, tout en contrôlant l'équivalence des différents groupes. L'usage d'un *test d'ancrage* implique toutefois des exigences supplémentaires (Angoff, 1971, p.578). Il faut tout d'abord que le test d'ancrage soit corrélé avec les tests à mettre en équivalence. Utiliser, par exemple, une épreuve de psychomotricité pour mettre en équivalence des tests de vocabulaire n'aurait guère de sens. De plus, le test d'ancrage doit représenter une tâche équivalente pour les différents groupes de sujets. Par ailleurs, bien que sa forme générale soit la même, l'équation permettant de déterminer les scores y^* est sensiblement plus complexe que pour les autres plans expérimentaux. Dans cette équation, la lettre " z " désigne les scores au test d'ancrage Z. L'indice " 1 " est utilisé pour les scores au test Z du groupe ayant passé le test X et l'indice " 2 " est utilisé pour les scores au test Z du groupe ayant passé le test Y.

$$y^* = a(x - c) + d \tag{7.11}$$

Détaillons les différentes composantes de cette formule :

$$a = \sqrt{\frac{s_Y^2 + b_{YZ_2}^2\left(s_Z^2 - s_{Z_2}^2\right)}{s_X^2 + b_{XZ_1}^2\left(s_Z^2 - s_{Z_1}^2\right)}} \tag{7.12}$$

s_X^2 est la variance des scores du premier groupe au test X

s_Y^2 est la variance des scores du second groupe au test Y

s_Z^2 est la variance des scores des deux groupes au test Z

$s_{Z_1}^2$ est la variance des scores du premier groupe au test Z

$s_{Z_2}^2$ est la variance des scores du second groupe au test Z

b_{XZ_1} est la pente de la droite de régression de X sur Z (groupe 1)

b_{YZ_2} est la pente de la droite de régression de Y sur Z (groupe 2)

$$c = \bar{X} + b_{XZ_1}(\bar{Z} - \bar{Z}_1) \qquad (7.13)$$

$\bar{X}$ est la moyenne des scores du premier groupe au test X

$\bar{Z}$ est la moyenne des scores des deux groupes au test Z

$\bar{Z}_1$ est la moyenne des scores du premier groupe au test Z

$$d = \bar{Y} + b_{YZ_2}(\bar{Z} - \bar{Z}_2) \qquad (7.14)$$

$\bar{Y}$ est la moyenne des scores du premier groupe au test Y

$\bar{Z}_2$ est la moyenne des scores du second groupe au test Z

Un exemple permettra d'illustrer cette dernière technique de mise en équivalence linéaire. Le tableau 1 présente les moyennes et les variances des scores de deux groupes aux deux tests à mettre en équivalence ainsi qu'à un test d'ancrage. Par ailleurs, la valeur de la pente de la droite de régression de X sur Z et celle de Y sur Z ont été calculées selon la procédure présentée dans le chapitre 2, §4. Ces valeurs sont les suivantes :

$b_{XZ_1} = 0{,}598$ et $b_{YZ_2} = 0{,}623$

À partir des différents résultats dont nous disposons, nous pouvons calculer les valeurs de a, c et d :

$$a = \sqrt{\frac{69{,}655 + 0{,}623^2\ (78{,}562 - 70{,}694)}{58{,}338 + 0{,}598^2\ (78{,}562 - (85{,}618))}} = 1{,}141$$

$$c = 23{,}097 + 0{,}773\ (12{,}949 - 12{,}464) = 23{,}472$$

$$d = 19{,}506 + 0{,}789\ (12{,}949 - 13{,}433) = 19{,}124$$

Grâce à ces valeurs, nous pouvons à présent calculer le score y équivalent à un score x donné. Par exemple, si $x = 18$, alors :

$$y^* = 1{,}141\ (18 - 23{,}472) + 12{,}880 \cong 13$$

Cette dernière valeur signifie qu'un score de 18 points sur le premier test est équivalent à un score de 13 points sur le second test.

Tableau 1 – Moyennes et variances des scores de deux groupes à deux tests et à un test d'ancrage

	Test 1	Test d'ancrage	Test 2
Groupe 1	$\bar{X} = 23{,}097$ $s_X^2 = 58{,}338$	$\bar{Z}_1 = 12{,}464$ $s_{Z_1}^2 = 85{,}618$	
Groupe 2		$\bar{Z}_2 = 13{,}433$ $s_{Z_2}^2 = 70{,}694$	$\bar{Y} = 19{,}506$ $s_Y^2 = 69{,}655$
Groupe 1 + Groupe 2		$\bar{Z} = 12{,}949$ $s_Z^2 = 78{,}562$	

2.3 LA MISE EN ÉQUIVALENCE ÉQUIPERCENTILE

La mise en équivalence linéaire repose sur des postulats particulièrement exigeants qu'il est difficile de satisfaire dans la pratique. Pour cette raison, une procédure s'appuyant sur des postulats plus faibles peut être préférée : la mise en équivalence équipercentile. Cette procédure est toutefois plus compliquée à mettre en oeuvre que la mise en équivalence linéaire. De plus, elle tend à produire des erreurs de mise en équivalence sensiblement plus importantes. Selon cette méthode, des scores au test X et au test Y sont considérés comme équivalents si leurs rangs centiles sont égaux. Concrètement, la procédure de mise en équivalence équipercentile est la suivante :

(1) Un groupe passe les deux tests dont les scores doivent être mis en équivalence (ou deux groupes tirés aléatoirement passe chacun un des tests).

(2) Pour chaque instrument, les équivalents centiles des différents scores brutes sont calculés.

(3) Pour chaque instrument, la relation entre les scores brutes et les rangs centiles est représentée graphiquement et les courbes sont lissées (voir § 1.3.5).

(4) Les scores brutes des deux instruments sont mis en équivalence via les percentiles correspondants. La figure 4 illustre cette dernière étape de la procédure. Les deux courbes lissées des tests X et Y sont tracées sur le même graphique. Il est alors aisé de mettre en relation les centiles et de déterminer les scores brutes qu'ils représentent dans chaque distribution des scores brutes. Ainsi, dans la figure 4, un score de 19 au test X est considéré comme équivalent à un score de 25 au test Y car ils correspondent tous les deux au centile 60.

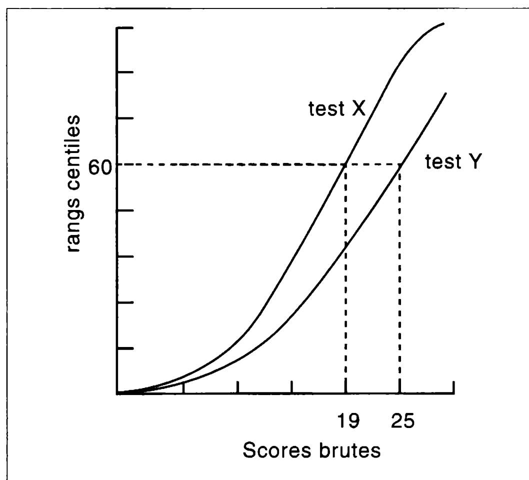


Figure 4 – Illustration graphique de la procédure de mise en équivalence équipercentile

3. Le calcul de scores seuil

3.1 Le concept de seuil de performance

Dans la section précédente, nous avons vu comment nous pouvions construire une échelle de mesure à partir des performances d'un échantillon représentatif de la population. Les graduations ainsi déterminées sont utiles pour comparer les performances d'un sujet à celles de la population à laquelle il appartient. Mais d'autres comparaisons intéressent également les praticiens. En particulier, ceux-ci peuvent désirer situer les performances d'un sujet par rapport à un niveau de performance souhaité. Dans ce cas, il est nécessaire de déterminer un score qui permettra de ranger les sujets en deux catégories : ceux qui atteignent le niveau souhaité et ceux qui ne l'atteignent pas. Un tel score de référence est qualifié de *score seuil.* Pour un même test, il est possible de fixer plusieurs scores seuil. Par exemple, pour un test de langue étrangère, nous pouvons déterminer différents scores correspondants chacun à un niveau nécessaire pour faire partie d'un groupe d'apprentissage.

Mais, le plus souvent, le praticien n'a besoin que d'un seul score seuil. C'est le cas lorsqu'il s'agit de décider si un élève atteint un niveau de maîtrise suffisant dans une matière donnée. Le score seuil est alors pris comme l'indicateur d'un niveau minimum de compétence. C'est également le cas lorsqu'il s'agit de décider si un candidat possède les compétences nécessaires pour occuper un poste de travail donné.

Comment déterminer un score seuil ? La réponse est loin d'être triviale. Les scores déterminés à partir d'impressions globales se sont, le plus souvent, révélés très peu valides. Ainsi, pour les examens scolaires, il est d'usage d'estimer le pourcentage d'erreur tolérable et, sur cette base, de déterminer le résultat minimum souhaité. Généralement, ce résultat est fixé à 50 ou 60% de réponses correctes. La pertinence de ces valeurs n'est guère fondée. Pour limiter au maximum les erreurs lors de prises de déci-

sion, il est nécessaire d'utiliser des méthodes plus rigoureuses pour déterminer les scores seuil. Comme nous le verrons plus loin, les méthodes actuelles restent imparfaites. Elles soulèvent plusieurs problèmes difficiles à résoudre. Nous soulignons dès à présent deux d'entre eux. Le premier provient du fait que la plupart des variables mesurées sont continues alors que nous souhaitons évaluer les compétences de manière dichotomique (compétent/non compétent). Le second problème découle de notre difficulté à définir les compétences minimales. Cette définition est souvent imprégnée par la subjectivité des juges et reste, par conséquent, relative.

Depuis le début des années 50 jusqu'aujourd'hui, un très grand nombre de méthodes ont été créées pour déterminer des seuils de performance les plus valides possible. Le lecteur intéressé en trouvera une large présentation dans l'ouvrage de V. de Landsheere (1988) " *Faire réussir, faire échouer* ". Dans le présent chapitre, nous ne détaillerons que les six méthodes qui semblent être aujourd'hui les plus utilisées (Kane, 1994). Nous pouvons les ranger en deux grandes catégories : (1) celles qui se basent sur le contenu du test et (2) celles qui se basent sur les performances des sujets.

3.2 MÉTHODES BASÉES SUR LE CONTENU DU TEST

Dans toutes ces méthodes, plusieurs juges passent en revue le contenu des items et, sur cette base, décident du niveau de performance suffisant pour réussir le test. Les diverses méthodes diffèrent par la technique utilisée pour atteindre cet objectif. Pour chacune d'elles, il existe plusieurs variantes que nous ne détaillerons pas ici.

La méthode de Nedelsky (1954) a été créée pour le cas des items à choix multiple. Pour chaque question, on demande aux juges de déterminer les choix de réponse qu'un sujet possédant une compétence minimale pourrait repérer comme incorrects. On peut dès lors déterminer la probabilité de répondre correctement à une question en choisissant une des alternatives restantes au hasard. Par exemple, si cinq choix de réponse sont proposés et qu'un sujet possédant une compétence minimale peut déterminer que trois de ces choix sont incorrects, le choix final de ce sujet ne se fera qu'entre les deux choix restant. Par conséquent, en répondant au hasard, ce sujet a une chance sur deux de choisir la réponse correcte. Son score probable est donc de 1/2 (ou de 0,50).

Une fois que l'on a déterminé par cette procédure le score probable à chaque item du test, on peut additionner ceux-ci pour obtenir le score total probable pour l'ensemble du test. Chaque juge ayant procédé de la sorte, il ne reste plus qu'à calculer la moyenne entre les scores probables déterminés par les différents juges. Cette valeur moyenne est le score le plus faible que devrait obtenir un sujet possédant une compétence minimale. Ce score définit ainsi un seuil entre les individus suffisamment compétents et les individus insuffisamment compétents.

Le tableau 2 présente une illustration de la détermination par un juge du score total probable pour un test de sept questions à choix multiple. Chaque choix de réponse est indiqué par une lettre. Le choix correct est indiqué en italique. La lettre est barrée si le juge estime qu'un sujet possédant une compétence minimale pourra déterminer que ce choix est incorrect. Nous pouvons constater que la somme des scores probables est égale à 3,11. En d'autres termes, si un sujet possède une compétence minimale, il

devrait obtenir au moins 3 points au test en question. La valeur de 3 points est ainsi le score seuil à ce test. La présente valeur n'a toutefois été déterminée que par un seul juge. Pour obtenir le score seuil de référence, il nous faudra encore calculer la moyenne des scores seuil déterminés par les différents juges.

Tableau 2 – Détermination du score seuil suivant la méthode de Nedelsky

Question	Réponses	Score probable
1	~~A~~ *B* ~~C~~ D ~~E~~	1/2 = 0,50
2	*A* B ~~C~~ D E	1/4 = 0,25
3	A ~~B~~ ~~C~~ D *E*	1/3 = 0,33
4	~~A~~ ~~B~~ ~~C~~ *D* E	1/2 = 0,50
5	~~A~~ *B* ~~C~~ ~~D~~ ~~E~~	1/1 = 1,00
6	*A* B C D E	1/5 = 0,20
7	A ~~B~~ ~~C~~ *D* E	1/3 = 0,33
Total		3,11

La méthode d'Angoff (1971) est vraisemblablement la plus utilisée aujourd'hui (Kane, 1994). Elle est utilisée pour toutes les formes de questions pour autant qu'elles soient cotées de manière dichotomique. Elle consiste à demander aux juges d'estimer la probabilité qu'un sujet possédant une compétence minimale aurait de réussir chacun des items du test. Cette méthode est simple dans son principe mais complexe dans sa réalisation. Il n'est en effet pas facile de traduire la compétence minimale en terme de probabilité. Pour cette raison, il est d'usage d'aider les juges dans leur tâche en leur proposant d'imaginer un groupe 100 sujets possédant une compétence minimale et d'estimer le nombre d'entre eux qui répondraient correctement à l'item en question. Chaque juge calcule un score seuil au test en additionnant les proportions de réussites aux différents items. Le tableau 3 illustre cette procédure. Le score seuil de référence est obtenu en calculant les moyennes des scores seuil déterminés par les différents juges.

Tableau 3 – Détermination du score seuil suivant la méthode d'Angoff

Question	% de réussites	Pourcentage/100
1	50	0,5
2	70	0,7
3	30	0,3
4	0	0,0
5	20	0,2
6	80	0,8
7	30	0,3
Total		2,8

La méthode d'Ebel (1972) est plus complexe que les deux précédentes car on demande aux juges de prendre en compte la pertinence et la difficulté des questions du test. Le travail des juges se déroule habituellement en deux temps. Lors de la première étape, chaque juge est invité à ranger les items dans un tableau en fonction de leur importance pour le programme d'apprentissage (ou pour l'activité professionnelle...) et en fonction de leur degré de difficulté. Le tableau 4 présente une illustration d'un tel classement.

Tableau 4 – Détermination du score seuil suivant la méthode d'Ebel

Importance et difficulté	Nombre d'items	Proportion de réussite	Score probable
Très important			
• très difficile	6	0,8	$6 \times 0,8 = 4,8$
• difficile	8	0,9	$8 \times 0,9 = 7,2$
• facile	6	1	$1 \times 0,6 = 0,6$
Important			
• très difficile	2	0,7	$2 \times 0,7 = 1,4$
• difficile	4	0,8	$4 \times 0,8 = 3,2$
• facile	4	0,9	$4 \times 0,9 = 3,6$
Peu important			
• très difficile	2	0,4	$2 \times 0,4 = 0,8$
• difficile	4	0,5	$4 \times 0,5 = 2,0$
• facile	0	-	-
total :			23,6

Une fois que tous les items ont été classés, on demande aux juges d'estimer le pourcentage de questions de chaque catégorie susceptible d'être réussies par un sujet possédant une compétence minimale. Les proportions ainsi déterminées sont alors multipliées par le nombre d'items correspondant. Par exemple, si le juge estime que le sujet sera capable de réussir 80% des items importants et difficiles, il faudra multiplier le nombre d'items de cette catégorie par 0,8. Nous obtenons ainsi le score probable d'un sujet ayant une compétence minimale dans la catégorie d'items en question. Le score seuil au test est obtenu en additionnant les scores probables aux différentes catégories d'items. Enfin, le score seuil de référence est déterminé en calculant la moyenne des scores seuil obtenus par les différents juges.

La méthode de Jaeger (1989) permet d'éviter le problème de la référence générale, et finalement assez abstraite, à un sujet possédant une compétence minimale. Cette méthode, présentée pour la première fois par Jaeger en 1978, est beaucoup plus contextualisée que les précédentes. Elle a par ailleurs la caractéristique d'être itérative. Elle consiste en une suite de réévaluations des mêmes items associées à la communication d'informations à propos de ces items.

Avant d'évaluer les questions, les juges sont invités à passer eux-mêmes le test afin de les familiariser avec les items qu'ils vont devoir évaluer. Il leur est ensuite demandé, pour chaque item, de répondre par oui ou par non à la question suivante (tableau 5) : " *Tous les sujets qui bénéficieront d'une décision favorable sur base des résultats du test [...] devraient-ils être capable de répondre correctement à cet item ?* " (Jaeger, 1989, p.494). Les juges doivent ainsi explicitement faire référence aux sujets réels qui seront évalués avec le test (par exemple, les élèves qui recevront un diplôme d'études secondaires sur base des résultats au test). Lorsque tous les items ont été évalués une première fois, les juges sont informés des estimations de leurs collègues et du pourcentage de sujets ayant réussi chacun de ces items lors d'un prétest. Les juges sont alors invités à réévaluer tous les items. On leur montre ensuite le pourcentage de sujets qui échoueraient si leurs évaluations des items étaient effectivement utilisées à des fins de classement. Après cela, les juges réévaluent une dernière fois l'ensemble des items. Le score seuil de référence est déterminé en calculant la médiane des scores seuil mis en évidence par chaque juge.

Tableau 5 – Détermination du score seuil suivant la méthode de Jaeger

Question	Doit-elle être réussie ?	Score attendu
1	oui	1
2	non	0
3	oui	1
4	oui	1
5	oui	1
6	oui	1
7	non	0
8	oui	1
9	oui	1
10	oui	1
11	oui	1
12	oui	1
Total		**10**

3.3 MÉTHODES BASÉES SUR LA PERFORMANCE DES SUJETS

Ces méthodes tentent de réduire la subjectivité dans la définition du score seuil en utilisant des données empiriques, en l'occurrence les résultats recueillis avec le test sur un échantillon de sujets. Pour que ces méthodes soient efficientes, il est nécessaire que les juges aient une expérience suffisante des sujets qui vont avoir à passer le test. Ils vont en effet classer ces sujets en fonction de leur niveau de compétence. La suite

de la procédure consistera en la passation du test par les sujets préalablement classés et en la détermination d'un score seuil sur base des scores observés.

Deux méthodes principales s'appuient sur les performances des sujets :

Dans la *méthode des groupes limites* (Livingstone & Zielsky, 1982), les juges doivent sélectionner au sein d'un groupe les sujets dont les compétences sont proches du niveau minimum attendu. Les sujets nettement plus faibles ou nettement plus forts sont donc écartés. Pour réaliser correctement cette tâche de sélection, on choisit habituellement comme juges des enseignants ou des formateurs qui connaissent bien les sujets de l'échantillon. Lorsque les sujets " limites " ont été sélectionnés, chacun de ceux-ci passe le test. Le score seuil est ensuite déterminé en calculant le score médian de la distribution des résultats des sujets " limites ". Le score seuil ainsi calculé n'a bien entendu de valeur que si les résultats sont bien groupés autour de la médiane.

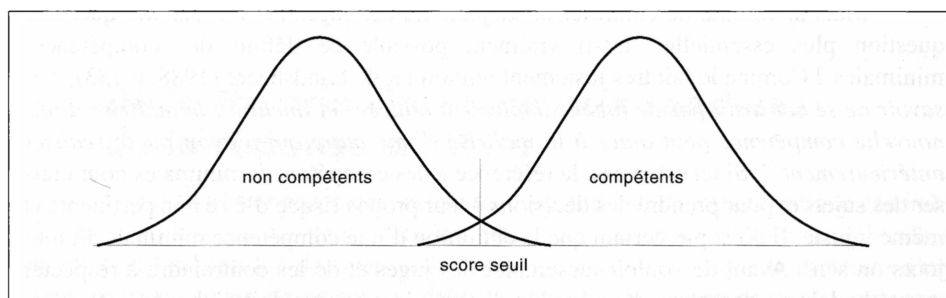


Figure 5 – Détermination du score seuil par la méthode des groupes contrastés

Dans la *méthode des groupes contrastés* (Livingstone & Zielsky, 1982), les juges sont invités à classer les sujets en deux groupes. D'un côté les sujets qu'ils jugent compétents et de l'autre ceux qu'ils jugent non compétents. Tous les sujets passent ensuite le test. Le plus souvent le score seuil est alors déterminé à l'aide d'une technique graphique (figure 5). La distribution des scores des sujets compétents et celle des sujets non compétents sont représentées simultanément. Le point d'intersection entre ces deux distributions représente le score seuil. La valeur correspondant au point d'intersection réduit au maximum les erreurs de classement des sujets. On minimise en effet le nombre de *faux négatifs*, c'est-à-dire de sujets classés comme non compétents sur base de leur résultat au test alors que les juges les avaient estimés compétents, et de *faux positifs*, c'est-à-dire de sujets classés comme compétents sur base de leur résultat au test alors que les juges les avaient estimés non compétents.

3.4 Validité des scores seuil

Toutes les méthodes que nous venons de présenter font appel à des jugements. Les racines de ceux-ci sont difficilement contrôlables. Cette part de subjectivité dans l'évaluation des items entraîne une certaine relativité des scores seuil déterminés selon ces méthodes. Deux groupes de juges évaluant le même ensemble d'items peuvent ainsi déterminer des scores seuil différents.

Les spécialistes de l'évaluation se sont donc attachés à réduire la subjectivité des juges afin d'améliorer la validité des scores seuil. On insiste à présent sur la néces-

sité d'utiliser un nombre suffisant de juges et de choisir ceux-ci de manière aléatoire. Il apparaît également nécessaire de soumettre les juges à un entraînement préalable et de leur donner des instructions claires à propos du contexte d'usage du test.

Un point qui a particulièrement retenu l'attention des chercheurs concerne la définition de la compétence minimale. Nous avons vu que la plupart des méthodes font appel à cette notion. Or les juges sont, au moins implicitement, influencés par les performances des sujets qu'ils connaissent lorsqu'ils se construisent une représentation de la compétence minimale. Celle-ci n'apparaît jamais *ex nihilo*. Plutôt que de tenter d'éliminer toute référence à des expériences antérieures (ce qui est impossible), il apparaît plus judicieux d'amener les juges à en prendre clairement conscience. C'est l'option prise par la méthode de Jaeger. La prise de conscience des références subjectives conduit chaque juge à un meilleur contrôle des facteurs qui influencent ses propres estimations.

Mais la volonté de contrôler la subjectivité des juges ne doit pas masquer une question plus essentielle : est-il vraiment possible de définir des compétences minimales ? Comme le fait très justement remarquer de Landsheere (1988, p.133), " *le savoir ne se construit pas de façon simplement additive et linéaire : la maîtrise d'une nouvelle compétence peut aider à la maîtrise d'une autre qui n'avait pu se réaliser antérieurement* ". Si tel est le cas, la référence à des compétences minimales pour classer des sujets et pour prendre des décisions à leur propos risque d'être non pertinente et même injuste. Il n'est pas certain que la définition d'une compétence minimale ait toujours un sens. Avant de vouloir rassembler des juges et de les contraindre à respecter un méthodologie complexe d'évaluation d'items, le praticien doit d'abord s'interroger sur la possibilité de définir des compétences minimales dans le domaine qu'il souhaite évaluer.

CHAPITRE 8

LES MODÈLES DE LA RÉPONSE À L'ITEM

1. De la théorie classique aux modèles de la réponse à l'item

La relativité des propriétés métriques des items est une caractéristique générale de l'analyse classique des items. Tous les indices que nous pouvons calculer dépendent en effet de l'échantillon de sujets utilisé. Ainsi, nous avons vu que la difficulté d'un item (sa valeur *p*) est classiquement définie comme la proportion de sujets qui répondent correctement à l'item. Par conséquent, si les sujets testés sont faibles, l'item sera considéré comme difficile. Par contre, si les sujets testés possèdent un niveau de compétence élevé, l'item sera considéré comme facile. Cette relativité de la valeur *p* a d'évidentes implications lors de l'application ultérieure des items. En effet, la capacité des sujets étant appréciée sur base d'une valeur *p* relative, le niveau de cette capacité sera elle-même relative. En d'autres termes, les caractéristiques de l'item sont dépendantes du groupe et les caractéristiques des sujets sont dépendantes des items.

Le problème de la relativité des propriétés métriques des items est particulièrement aigu dans le cas d'une banque d'items (c'est-à-dire un vaste ensemble d'items d'où l'on puise pour construire des tests). En effet, les items qui la composent ne sont habituellement pas analysés avec le même groupe de sujets. À chaque création d'un nouvel ensemble d'items, une étude empirique est réalisée. Les sujets utilisés pour cette analyse changent, mais aussi le moment d'application. Ce dernier point est crucial lorsqu'il s'agit d'items évaluant des acquis scolaires. En effet, les élèves testés en octobre seront forcément plus faibles que ceux qui seront testés en mai car les premiers seront en début d'apprentissage alors que les seconds auront bénéficié d'une longue période d'exercice. Par conséquent, les items qui composeront la banque posséderont des caractéristiques métriques qui ne seront pas comparables. Comment dès lors composer un test avec de tels items ?

Pour construire une banque d'items efficiente, l'idéal semble donc de pouvoir obtenir des caractéristiques d'items qui soient indépendantes du groupe de sujets qui ont permis de les calculer. Plus largement, pour de nombreuses applications en psycho-

logie et en éducation, il apparaît très utile de pouvoir construire des échelles de mesure indépendantes d'un groupe de référence, c'est-à-dire des échelles absolues. En effet, la signification du score total à de telles échelles n'est plus relative aux caractéristiques d'un groupe de référence. Une tentative pour développer une échelle de ce type a été faite au début des années 50 par Guttman dans le but de mesurer des attitudes. Sur une échelle de Guttman, un sujet qui répond par l'affirmative à une question reflétant une attitude très marquée doit également répondre par l'affirmative à une question reflétant un degré moins marqué de la même attitude. Et réciproquement. Une illustration d'une échelle de Guttman dans le cadre du modèle piagétien du développement cognitif est présentée dans le chapitre 5 (§5.4).

Lorsque nous sommes en présence d'une échelle parfaite de Guttman, la seule connaissance du score total d'un sujet nous permet de déterminer exactement les scores qu'il a obtenus à chacun des items. Un test qui rencontre les exigences du modèle de Guttman peut être qualifié d'homogène, d'unidimensionnel et de fiable (Angoff, 1971, p.529). En effet, tous ses items évaluent un seul et même trait psychologique et permettent de situer de manière très précise un sujet sur le continuum mesuré. Dans la réalité, de tels tests sont rares car les exigences du modèle de Guttman sont très difficiles à satisfaire. La performance des sujets doit être entièrement déterminée par leur seule position sur le continuum mesuré. Aucune autre variable ne peut influencer leur performance. Pour cette raison, le modèle de Guttman est qualifié de strictement déterministe (Matalon, 1965, p.33).

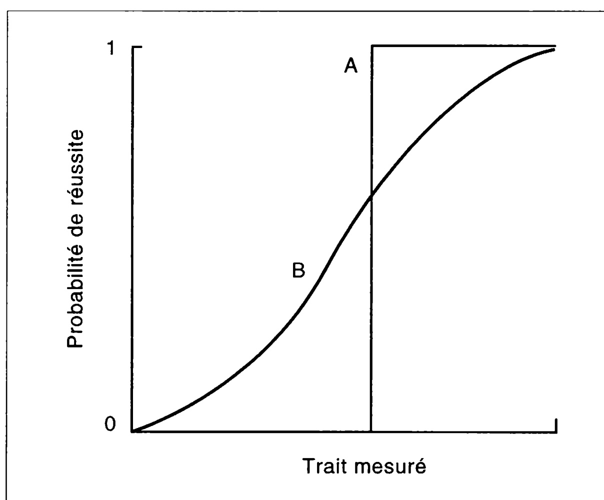


Figure 1 – Modèle déterministe (A) et modèle probabiliste (B) de la réponse à un item

La figure 1 illustre de manière graphique le modèle déterministe proposé par Guttman. Le trait mesuré est indiqué en abscisse et la probabilité de réussir l'item est indiquée en ordonnée. Selon le modèle de Guttman, un item a une probabilité nulle d'être réussi au-dessous d'un certain niveau de capacité. Par contre, à partir de ce niveau et au-dessus de celui-ci, la réussite de l'item est certaine. Ce passage d'une probabilité nulle à une probabilité égale à 1 est représenté par une droite perpendiculaire à l'abscisse (A). Du fait des inévitables erreurs de mesure, on comprend aisément qu'il

soit peu vraisemblable de rencontrer une telle situation dans la réalité. Pour cette raison, des modèles probabilistes ont aujourd'hui remplacé le modèle déterministe de Guttman. Dans ce cas, plus le sujet se situe à un niveau élevé sur le trait, plus sa probabilité de réussir l'item augmente. La courbe B illustre cette élévation progressive de la probabilité de réussite en fonction du degré d'habileté du sujet.

Les modèles probabilistes s'appuient sur le postulat qu'une réponse correcte à l'item est déterminée par : le trait mesuré, la difficulté de l'item et la discrimination de l'item. En d'autres termes, la probabilité de réussite d'un item est une fonction de la caractéristique psychologique du sujet (le trait mesuré) et des propriétés métriques de l'item (sa difficulté et sa discrimination). Les psychométriciens ont proposé divers modèles de relation fonctionnelle entre l'item et le trait mesuré. Tous ces modèles partagent le postulat que tous les items d'un test mesurent une même caractéristique psychologique mais que le patron des réponses à ces items peut être affecté par des erreurs aléatoires. Tous ces modèles ont également pour objectif de permettre, d'une part, une estimation des propriétés métriques des items invariantes au travers des populations et, d'autre part, une estimation des traits psychologiques indépendante des items utilisés pour les mesurer.

Les modèles probabilistes sont aujourd'hui rassemblés dans la catégorie générale des Modèles de la Réponse à l'Item (MRI). Nous présentons les plus importants de ces modèles dans la section suivante.

2. La fonction caractéristique de l'item

Le postulat de base des MRI est que la performance d'un sujet à un item peut être expliquée par un facteur appelé *trait latent*. Ce dernier terme a ici un sens très général. En effet, le trait latent peut être un trait de personnalité, une aptitude cognitive, une compétence scolaire...

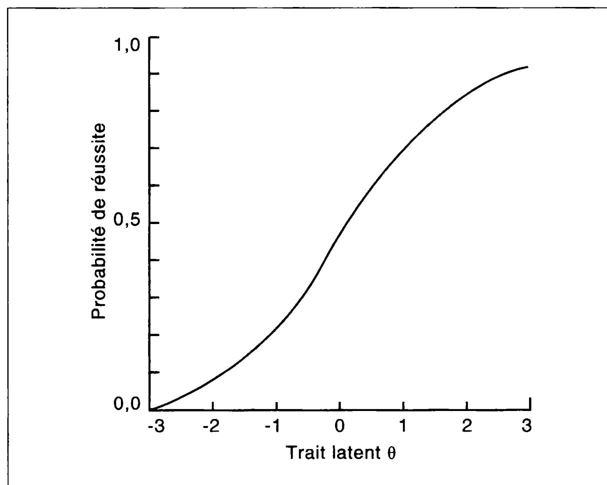


Figure 2 – La courbe caractéristique de l'item

La relation entre les performances à l'item et le trait latent peut être décrite au moyen d'une fonction appelée *fonction caractéristique de l'item* ou, plus couramment, *courbe caractéristique de l'item* (CCI). Cette courbe prend habituellement la forme d'un S plus ou moins allongé. La figure 2 présente un exemple de CCI. Le trait latent apparaît en abscisse et est traditionnellement représenté par la lettre grecque θ *(thêta)*. Le trait latent étant distribué normalement au sein de la population, les graduations de l'abscisse correspondent aux valeurs de la distribution de *z*. Le niveau moyen est donc représenté par la valeur 0 et la distance d'un écart type par rapport à cette moyenne est représenté par + ou - 1. Sur la figure 2, nous avons indiqué des graduations allant de -3 à +3. Cet intervalle inclut 99,8% des sujets. Si nous souhaitons inclure une proportion encore plus grande de sujets, nous pouvons bien entendu faire débuter la graduation de l'abscisse par une valeur inférieure à -3.

Sur le même graphique, la probabilité de donner une réponse correcte à l'item apparaît en ordonnée. Les valeurs de *y* s'étendent de 0 à 1. Plus un sujet se situe à un niveau élevé sur le trait latent, plus sa probabilité de répondre correctement à l'item est grande. Et réciproquement. La probabilité de réussite dépend également de la difficulté de l'item. A valeurs égales de θ, la valeur de *y* augmente ou diminue selon ce niveau de difficulté.

Par convention, la valeur qui représente la difficulté d'un item est égale à la valeur de θ pour laquelle la probabilité de donner une réponse correcte est de 0,5. Sur la figure 3, l'item correspondant à la courbe A possède une difficulté égale à 1 et l'item correspondant à la courbe B possède une difficulté égale à 2. Le niveau de difficulté d'un item représente un premier paramètre permettant de décrire la CCI de cet item.

Mais la difficulté n'est pas le seul paramètre en jeu. Un second paramètre important est la pouvoir discriminatif de l'item. La discrimination d'un item est représentée par la pente de la CCI. Celle-ci peut être plus ou moins inclinée. Plus la pente est abrupte, plus l'item est discriminatif. Et inversement. Sur la figure 3, les items représentés par les courbes A et B sont très discriminatifs. Par contre, la courbe C est caractéristique d'un item modérément discriminatif.

Les premiers travaux à propos de la fonction qui relie un item au trait latent remontent au début des années 50. Les contributions de Lord (1953a; 1953b) sont particulièrement importantes. Mais les développements théoriques et les applications des MRI ont été particulièrement stimulés par la publication en 1960 d'un article du mathématicien danois Georg Rasch : *« Probalistic models for some intelligence and attainment tests »*. Rasch semble avoir été le premier à avoir utilisé une fonction logistique pour analyser des données dans le cadre de la construction d'un test psychologique (Wright & Stone, 1979, p.X). Le modèle proposé par Rasch est le plus simple des MRI. Il s'appuie en effet sur le postulat que tous les items discriminent également. Par conséquent, le seul paramètre à estimer concerne la difficulté des items.

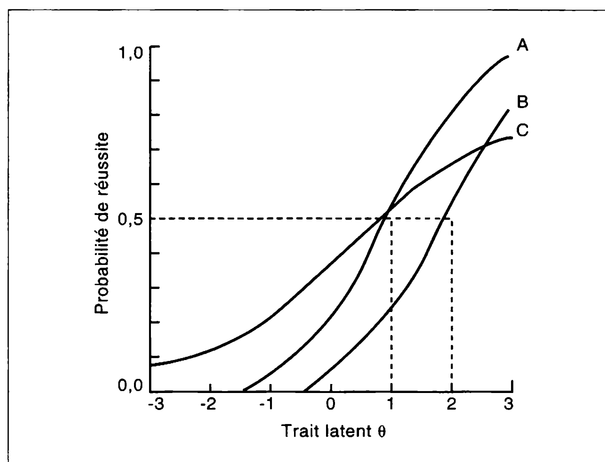


Figure 3 – CCI pour trois items dont la difficulté et le pouvoir discriminatif diffèrent

Le modèle à un paramètre, souvent appelé *« modèle de Rasch »*, est aujourd'hui le plus simple à utiliser des MRI. Selon ce modèle, la probabilité de réussite à un item peut être estimée par la formule suivante :

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad (8.1)$$

$P_i(\theta)$ = probabilité qu'un sujet quelconque, possédant une aptitude, réponde correctement à l'item i,

b_i = paramètre de difficulté de l'item i,

e = constante de Neper, valeur qui correspond au nombre irrationnel 2,718281...

Le modèle de Rasch est particulièrement exigeant puisqu'il postule que tous les items d'un test possèdent la même discrimination. Cette exigence peut être rencontrée lorsque les items sont très semblables comme, par exemple, dans les tests d'acquis scolaires focalisés sur un domaine précis. Mais, dans beaucoup d'autres cas, cette exigence n'est pas aisée à satisfaire. Pour cette raison, un modèle qui prend en compte la diversité de la puissance discriminative des items est de plus en plus utilisé. Le modèle logistique à deux paramètres (difficulté et discrimination) a été développé par Birnbaum (1968). Suivant ce modèle, la probabilité de réussite à un item peut être estimée par la formule suivante, qui est une extension de l'équation 8.1 :

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (8.2)$$

a_i = paramètre de discrimination de l'item i ; il est proportionnel à la pente de la courbe au point b_i,

D = facteur d'échelonnement (*scaling factor*); il prend une valeur constante égale à 1,7.

La valeur de a_i se situe habituellement entre 0 et 2. Lorsque cette valeur est négative, cela signifie que la probabilité de réussir l'item diminue en fonction de l'habileté du sujet. Une telle situation n'a guère de sens d'un point de vue psychologique. Par conséquent, un item présentant un paramètre de discrimination négatif est habituellement éliminé ou, au minimum, révisé. Par ailleurs, il est rare de rencontrer des items dont la valeur a_i est supérieure à 2. Une telle valeur indique une pente particulièrement raide. Du fait des inévitables erreurs de mesure, il est peu probable d'observer une discrimination plus marquée.

La majorité des recherches sur les MRI ont été réalisées avec des items à choix multiple ou dont les réponses étaient du type « vrai/faux ». Lorsque l'on utilise de telles modalités de réponse, le risque existe que des sujets ne possédant aucune habileté réussissent malgré tout un item en répondant au hasard. Dans ce cas, l'asymptote la plus basse de la CCI est nettement supérieure à zéro (voir la courbe C sur la figure 3). Pour faire face à une telle éventualité, il a été proposé d'inclure un troisième paramètre dans l'équation 8.2 : le paramètre de *« pseudo-chance »*. Cette dénomination peut étonner. En fait, comme les valeurs de ce paramètre sont habituellement inférieures à celles auxquelles correspondrait un choix totalement aléatoire, on considère qu'il n'est pas exact de l'appeler *« paramètre de chance »* (Hambleton, Swaminathan & Rogers, 1991, p.17). L'équation suivante correspond au modèle à trois paramètres :

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (8.3)$$

c_i = paramètre de pseudo-chance.

L'avantage majeur des MRI que nous venons de présenter est de nous permettre de déterminer les paramètres caractéristiques d'un item (difficulté, discrimination et pseudo-chance) indépendamment des caractéristiques des sujets qui ont permis de les estimer. Cette propriété d'invariance des paramètres rend possible l'usage d'une banque d'items. En effet, le praticien désireux de constituer un test peut sélectionner des items qui n'ont pas été analysés avec les mêmes échantillons de sujets. Une multitude de combinaisons d'items est dès lors possible. Ceci est particulièrement intéressant lorsque l'on veut retester des sujets en évitant l'effet de répétition des mêmes items. Un autre avantage est de pouvoir comparer des sujets qui ont passé des ensembles d'items différents comme, par exemple, lors de testing adaptatifs (voir §5.2.). En effet, les mesures obtenues sont indépendantes des items particuliers qui ont été utilisés.

L'invariance des paramètres peut étonner les praticiens habitués à la relativité des analyses classiques. Une comparaison simple peut aider à comprendre cette propriété (Hambleton, Swaminathan & Rogers, 1991, p.19). Dans le modèle de régression linéaire, la relation entre une variable X et une variable Y est représentée par une droite de régression. Cette droite est décrite par une équation dont les paramètres sont la pente et l'ordonnée à l'origine. Ces paramètres sont estimés à partir d'un ensemble restreint de valeurs de X. Si le modèle de régression linéaire est adéquat, tout autre ensemble de valeurs de X devrait aboutir à la mise en évidence des mêmes paramètres. Il est logique que cette propriété d'invariance des paramètres s'applique aussi aux MRI qui peuvent être vus comme des modèles de régression non linéaire.

Bien que les MRI apparaissent comme très séduisants au premier abord, le praticien ne doit pas perdre de vue qu'ils reposent sur des postulats très forts. Avant d'utiliser les MRI, il est par conséquent nécessaire de vérifier si certaines exigences sont satisfaites au niveau des items et de la réalité qu'ils mesurent. Les deux postulats sur lesquels s'appuient les MRI que nous venons de présenter sont : l'unidimensionnalité et l'indépendance locale.

L'*exigence d'unidimensionnalité* signifie que tous les items d'un test doivent mesurer un seul et même trait. Dans la pratique, ce critère n'est jamais parfaitement rencontré du fait des inévitables erreurs de mesure et de la complexité des traits mesurés. En fait, *« il est raisonnable de considérer qu'un ensemble de n tests ou de n items dichotomiques est unidimensionnel si et seulement s'il s'ajuste à un modèle factoriel non linéaire avec un seul facteur commun. Dans le cas de tests, il est généralement correct de considérer que la régression de chaque score au test sur le facteur commun est linéaire. [...]. Dans le cas de données dichotomiques ce postulat n'est jamais correct car la régression de l'item sur le facteur est [...] une probabilité conditionnelle dont les bornes sont zéro et un »* (McDonald, 1981, p. 104-105). Lorsque cette exigence d'unidimensionnalité ne peut être rencontrée, les MRI ne peuvent être utilisés. Actuellement, des MRI multidimensionnels sont en cours de développement. Mais ils sont compliqués à utiliser et comportent de nombreuses contraintes, par exemple en terme de taille d'échantillon.

L'*exigence d'indépendance locale* signifie quant à elle que le trait qui fait l'objet de l'évaluation doit être le seul facteur qui détermine la variabilité des réponses aux items d'un test. Une fois que le trait mesuré a été pris en compte, aucune relation ne doit exister entre les réponses d'un sujet aux différents items. Si, par exemple, les consignes d'un test donnent des indices permettant de répondre plus facilement à certains items, l'exigence d'indépendance locale n'est plus respectée. Certains sujets peuvent en effet remarquer cet indice et d'autres pas. Par conséquent, le score au test ne dépendra pas seulement du trait que l'on veut mesurer mais également de la capacité à repérer certains indices utiles. L'exigence d'indépendance locale n'est pas non plus satisfaite si, par exemple, dans un test de mathématique, certains items font appel à des connaissances en géographie. En effet, la réussite de ces items n'est pas déterminée par le seul trait latent que nous souhaitons mesurer. Les sujets qui possèdent de bonnes connaissances en géographie auront une probabilité plus élevée que les autres sujets d'obtenir un score élevé au test de mathématique.

Lorsque l'exigence d'unidimensionnalité est satisfaite, l'exigence d'indépendance locale l'est aussi. L'inverse n'est cependant pas vrai. Nous pouvons observer un espace latent multidimensionnel et en même temps une indépendance locale des items du test. Cette situation se produit lorsqu'un second facteur influence tous les items de manière égale. Par exemple, dans un test de mathématique, les sujets peuvent avoir à lire de courts énoncés. Si tous les sujets lisent couramment, ce facteur ne les différenciera pas. Il y aura alors indépendance locale : les sujets se distingueront selon leur seule compétence en mathématique. Pourtant, le test ne pourra être considéré comme unidimensionnel puisque les performances seront sous-tendues par au moins deux facteurs.

Par ailleurs, l'utilisation des MRI implique certaines contraintes méthodologiques. Il est évident que l'utilisation pratique de ces modèles est plus complexe que

celle des techniques issues de la théorie classique des tests. Elle demande aux praticiens de sérieuses compétences théoriques et des outils informatiques puissants. Ceci limite certainement le champ d'application des MRI. Parmi les contraintes méthodologiques, les deux plus importantes concernent l'estimation des paramètres et la procédure de liaison.

La première contrainte méthodologique concerne *l'estimation des paramètres.* La procédure d'estimation des paramètres est souvent appelée « *calibration* » des items. Cette procédure est relativement complexe et, bien que certains aient proposé des procédures manuelles de calcul (Wright & Stone, 1979, pp.28-44), le recours à l'ordinateur est presque toujours nécessaire. Ce recours est d'autant plus nécessaire que le nombre d'items et de sujets nécessaires pour une calibration précise des paramètres est assez élevé (Hulin, Lissak & Drasgow, 1982). Il existe aujourd'hui sur le marché de nombreux programmes permettant de calibrer les items pour les MRI à un, deux ou trois paramètres. Hambleton, Swaminathan et Rogers (1991, pp.46-50) en font une intéressante présentation critique.

Remarquons ici que la création de banques d'items découle en partie des contraintes d'estimation des paramètres. Comme le souligne Van Der Linden (1986, p.330), « *une banque d'items sans MRI est irréalisable. Mais il est également vrai que le potentiel des MRI peut seulement se réaliser en combinaison avec une banque d'items* ». En effet, souvent, la détermination des paramètres ne se fait pas en une fois mais, au contraire, par approximations successives. À chaque utilisation de l'item, les résultats sont incorporés à l'ensemble des résultats antérieurs, ce qui permet d'améliorer la calibration de l'item en question. Il existe ainsi une relation réciproque et dynamique entre une banque d'items et les tests construits à partir d'elle.

La seconde contrainte méthodologique concerne *la procédure de liaison (linking).* Nous avons souligné plus haut que, dans le cadre des MRI, les paramètres des items étaient invariants. Dans la pratique, ce n'est pas tout à fait exact. En effet, lors la procédure d'estimation des paramètres, le point zéro de l'échelle de difficulté des items est arbitrairement centré sur la moyenne des estimations de θ pour le groupe de sujets utilisé pour l'analyse. La position du zéro variant selon les groupes, les paramètres obtenus sont ipso facto relatifs. Ce problème peut heureusement être résolu assez aisément car les paramètres d'un item sont effectivement invariants compte tenu d'une transformation linéaire qui permet de les placer tous sur une même échelle. En d'autres termes, lors du calibrage des items, il est nécessaire de déterminer une constante qui permettra de transformer les valeurs obtenues et de les ajuster à l'échelle de référence. Pour réaliser cette procédure de liaison, plusieurs techniques sont possibles (Vale, 1986) : placer des items communs dans les différents ensembles d'items ou utiliser des sujets communs lors des passations des divers groupes d'items, ou les deux à la fois.

3. L'estimation des paramètres

L'estimation des paramètres des items est une opération cruciale. En effet, c'est la qualité de cette estimation qui donne son sens à l'utilisation des MRI. Si l'estimation est mauvaise, les paramètres seront instables d'un échantillon de sujets à l'autre. Par conséquent l'intérêt des MRI sera perdu puisque nous n'obtiendrons pas d'invariance des paramètres. De nombreuses procédures d'estimation des paramètres ont été propo-

sées depuis les premiers travaux de Rasch. La plus utilisée aujourd'hui est certainement la *méthode du maximum de vraisemblance marginale (marginal maximum likelihood).* Elle est en effet utilisée par les programmes informatiques les plus courants comme BILOG (Mislevy & Bock, 1990) ou XCALIBRE (Assessment Systems Corporation, 1994).

Lorsque nous voulons estimer les paramètres d'un ensemble d'items, les réponses des sujets sont les seules informations dont nous disposons. Nous sommes alors contraints d'estimer en même temps le trait θ des sujets et les paramètres des items. La réalisation de cette double estimation est loin d'être évidente. Pour en comprendre la logique, il est nécessaire de partir d'une situation plus simple où les paramètres sont connus et où seule le trait θ des sujets doit être estimé sur base de leur patron de résultats. La probabilité qu'un sujet possédant une capacité θ obtienne une réponse Uj (Uj = 1 si la réponse est correcte et 0 si elle est fausse) se note $P(U_j/\theta)$. En vertu du postulat d'indépendance locale, la probabilité d'observer un patron de réponse à un ensemble de *n* items est égale au produit des probabilités de réponse à chacun de ces items :

$$P(U_1, U_2, \ldots, U_n|\theta) = \prod P(U_j|\theta) \quad (8.4)$$

La formule 8.4 peut également s'écrire :

$$P(U_1, U_2, \ldots, U_n|\theta) = \prod P_j^{U_j} Q_j^{1-U_j} \quad (8.5)$$

$$P_j = P(U_j|\theta)$$

$$Q_j = 1 - P(U_j|\theta)$$

Si nous plaçons dans cette formule le patron de réponses effectivement observé, celle-ci ne peut plus être interprétée de manière probabiliste. La fonction que nous obtenons alors est appelée la *fonction de vraisemblance :*

$$L(u_1, u_2, \ldots, u_n|\theta) = \prod P_j^{U_j} Q_j^{1-U_j} \quad (8.6)$$

L'estimation du trait θ d'un sujet consiste dès lors à calculer la valeur de θ qui maximise la fonction de vraisemblance 8.6. Pour trouver la valeur maximum de cette fonction, on utilise une procédure par approximations successives dont la plus courante est celle de Newton-Raphson. Aucune valeur finie ne peut toutefois être trouvée lorsque les réponses d'un sujet sont toutes correctes ou toutes erronées. Dans ce cas, l'estimation qui maximise la fonction de vraisemblance est $\theta = +\infty$ ou $\theta = -\infty$.

Lorsque nous ne connaissons ni les valeurs de θ ni les valeurs des paramètres, la situation est encore plus complexe car nous devons considérer en même temps l'ensemble des *n* items du test et les patrons de réponses des *N* sujets qui ont répondu à ces items. Dans ce cas, la fonction de vraisemblance s'écrit :

$$L(u_1, u_2, \ldots, u_j, \ldots, u_N|\theta, a, b, c) = \prod_{i=1}^{N} \prod_{j=1}^{n} P_{ij}^{U_{ij}} Q_{ij}^{1-U_{ij}} \quad (8.7)$$

Dans la formule 8.7, nous avons envisagé le cas où trois paramètres doivent être estimés (a, b et c). La même formule peut bien entendu être adaptée pour les cas où seulement un ou deux paramètres doivent être estimés.

Pour déterminer les valeurs de θ et les paramètres des items qui maximisent la fonction de vraisemblance 8.7, deux méthodes peuvent être suivies : (1) l'estimation du maximum de vraisemblance conjointe (ou conditionnelle); (2) l'estimation du maximum de vraisemblance marginale (ou inconditionnelle).

La méthode du maximum de vraisemblance conjointe présente certaines faiblesses. La première est que les paramètres des items échoués ou réussis par tous les sujets ne peuvent être estimés. Il en est de même pour les valeurs de θ lorsque les sujets n'obtiennent que des réponses correctes ou des réponses fausses. Pour cette raison, les logiciels qui utilisent la procédure du maximum de vraisemblance conjointe éliminent d'emblée les sujets et les items dont les scores sont uniquement 1 ou uniquement 0. Une seconde faiblesse de la méthode apparaît avec les modèles à deux et trois paramètres pour lesquels les estimations sont instables si l'on n'utilise pas un très grand nombre de sujets et d'items. Pour cette dernière raison, la méthode du maximum de vraisemblance conjointe n'est plus aujourd'hui utilisée que dans les logiciels qui réalisent des analyses selon le modèle de Rasch (par exemple, RASCAL, Assessment Systems Corporation, 1992). Dans les autres cas, on lui préfère la méthode du maximum de vraisemblance marginale. Cette méthode est beaucoup plus coûteuse en calculs que la précédente mais permet d'obtenir des estimations plus stables pour les modèles à deux ou trois paramètres.

Quelle que soit la procédure utilisée, les caractéristiques de l'échantillon de sujets jouent un grand rôle dans la qualité de l'estimation des paramètres des items. En particulier, *« un échantillon homogène de sujets entraînera des estimations instables des paramètres du modèle »* (Hambleton, 1994, p.541). Dans le but de garantir cette hétérogénéité et de réduire l'impact des erreurs de mesure, un échantillon de sujets suffisamment important est nécessaire. Hulin & al. (1982) ont évalué la taille optimale de cet échantillon pour les modèles à deux et à trois paramètres. Pour ce faire, ils ont généré des données simulées pour des échantillons de 200, 500, 1000 et 2000 sujets à des tests de 15, 30 et 60 items. Ces données ont été analysées avec le logiciel LOGIST. Il apparaît que, pour le modèle à deux paramètres, un test de 30 items et un échantillon de 500 sujets permettent d'obtenir des estimations de paramètres relativement stables. Avec le modèle à trois paramètres, le même objectif peut être atteint avec un test de 60 items et un échantillon de 1000 sujets. Dans le cas du modèle à un paramètre, Wright & Stone (1979) recommandent d'utiliser un minimum de 20 items et un échantillon de 200 sujets pour obtenir une estimation satisfaisante. Excepté dans ce dernier cas, nous pouvons nous rendre compte que l'estimation des paramètres est une procédure relativement coûteuse puisqu'elle impose une importante récolte de données. Cette exigence constitue une réelle limite pour l'application des MRI les plus sophistiqués.

Dans la pratique, les données utilisées pour l'estimation des paramètres ne s'ajustent jamais parfaitement au modèle choisi. Divers tests statistiques ont donc été mis au point pour évaluer le degré d'ajustement au modèle. Lorsque cette adéquation est faible, nous ne pouvons obtenir des estimations invariantes de paramètres. Les items mal ajustés devront par conséquent être écartés. Toutefois, il y a lieu d'être prudent avant de rejeter un item car les tests d'ajustement se sont révélés très sensibles à la

taille des échantillons de sujets. Lorsque cet échantillon est petit, des problèmes d'ajustement relativement importants peuvent ne pas être détectés. Par contre, lorsque l'échantillon est très grand, des problèmes d'ajustement mineurs risquent de conduire au rejet des items incriminés. Dans le tableau 2, nous reprenons, à titre d'illustration, les résultats présentées à ce propos par Hambleton & Murray (1983). Les données ont été analysées à l'aide du programme BICAL (Wright & Stone, 1979). Ce programme permet de réaliser une analyse des items selon le modèle de Rasch à l'aide de la procédure du maximum de vraisemblance conjointe. Il calcule également la valeur de *t* pour détecter les items mal ajustés au modèle au seuil de 0,01 et 0,05. Comme nous pouvons nous en rendre compte à la lecture du tableau 1, le nombre d'items mal ajustés que détecte le programme BICAL varie sensiblement selon la taille de l'échantillon de sujets.

Tableau 1 – Nombre d'items mal ajustés en fonction de la taille de l'échantillon (Hambleton & Murray, 1983)

Taille de l'échantillon	Nombre d'items mal ajustés (N=50)	
	$p < 0{,}05$	$p < 0{,}01$
150	20	5
300	25	17
600	30	18
1200	38	28
6000	42	38

Le problème que nous venons de souligner n'est pas spécifique à un logiciel informatique ni à un test d'ajustement. Il s'agit, au contraire, d'un problème tout à fait général. Par conséquent, Hambleton & Swaminathan (1985) suggèrent de ne pas focaliser son attention sur les tests d'ajustement. Ils recommandent de mener une investigation plus large concernant l'adéquation entre le modèle et les données. Les conclusions sur cette question doivent en effet découler de la convergence d'un ensemble d'indices. Trois catégories d'indices devraient retenir l'attention des praticiens (pour une présentation plus détaillée, voir Hambleton & Swaminathan, 1985, pp. 155-167 et Hambleton & al., 1991, pp. 55-74) :

1. *Les informations concernant la validité des postulats du modèle utilisé pour analyser les données.* Par exemple, les résultats des analyses concernant l'unidimensionnalité de l'ensemble des items font partie de ces informations.
2. *Les informations concernant les propriétés attendues sur base du modèle.* Ainsi, on peut évaluer si la propriété d'invariance des paramètres est confirmée en comparant les paramètres obtenus sur plusieurs échantillons de sujets de la population.
3. *Les informations concernant les prédictions réalisées sur base du modèle.* Par exemple, nous pouvons comparer la différence entre les performances effectives d'un groupe de sujets à un item et celles qui ont pu être prédites sur base du niveau d'aptitude (la valeur de θ) de ces mêmes sujets.

4. La fonction d'information de l'item et du test

Les paramètres d'un item peuvent nous renseigner à propos de la quantité d'information que nous procure cet item. L'information donnée par un item est maximale lorsque sa difficulté correspond au niveau d'aptitude du sujet évalué. Ainsi, un item de difficulté moyenne sera le plus informatif à propos des sujets dont le niveau d'aptitude est proche de la moyenne. Par contre, il ne nous apprendra pas grand chose à propos des sujets faibles (ils échouent tous), ni des sujets brillants (ils réussissent tous). Par ailleurs, l'information sera d'autant plus grande que la discrimination de l'item est élevée. Inversement, un item peu discriminatif nous donnera peu d'information. Enfin, plus le risque de réponse aléatoire est faible, plus l'item sera informatif. Lorsque nous avons affaire à des items dichotomiques, l'information que nous procure un item à propos d'un trait donné peut être évaluée à l'aide de la formule suivante :

$$I_i(\theta) = \frac{[P'_i(\theta)]^2}{P_i(\theta)\,Q_i(\theta)} \quad (8.8)$$

$I_i(\theta)$ = la fonction d'information d'un item i à propos du trait θ,

$P_i(\theta)$ = la fonction caractéristique de l'item (voir équations 8.1, 8.2 et 8.3),

$P'_i(\theta)$ = la dérivée première de $P_i(\theta)$,

$Q_i(\theta) = 1 - P_i(\theta)$.

Connaissant les fonctions d'information des items, nous pouvons calculer l'information que nous donne un test en fonction de θ. Du fait de l'indépendance locale des items, la fonction d'information d'un test est égale à la somme des fonctions d'information des items qui composent ce test :

$$I(\theta) = \sum I_i(\theta) \quad (8.9)$$

Cette formule est particulièrement utile pour les constructeurs de tests. Comme les items contribuent de manière indépendante à l'information donnée par le test dans son ensemble, il est relativement aisé de comparer différentes combinaisons d'items afin d'obtenir le test qui nous donne le maximum d'information dans la zone d'aptitude souhaitée. La figure 4 présente de manière graphique les fonctions d'information de deux tests. Les deux fonctions ont été obtenues à l'aide du logiciel RASCAL 3.5 (Assessment Systems Corporation, 1992) qui permet de réaliser une analyse d'item selon le modèle de Rasch en utilisant la procédure du maximum de vraisemblance conjointe. Par conséquent, l'information que nous procure chaque test est relative aux seuls paramètres de difficulté des items. Le premier test (A) comprend 46 items. Nous pouvons constater qu'il nous permet d'obtenir un niveau élevé d'information sur toute l'étendue du trait mesuré. Le second test (B) comprend 22 items. Par rapport au test A, l'information qu'il nous procure dépend nettement plus du niveau d'aptitude du sujet évalué. En fait, le test B comprend beaucoup d'items difficiles et même très difficiles. Par contre, il manque d'items de difficulté moyenne et inférieure à la moyenne. Remarquons ici que la fonction d'information d'un test ne doit pas a priori correspon-

dre à un modèle particulier La qualité de la courbe d'information d'un test dépend avant tout des besoins du praticien. Si le but de celui-ci est d'évaluer uniquement des sujets doués, un test informatif dans la seule zone supérieure du trait fera certainement l'affaire. Si, par contre, le praticien souhaite pouvoir évaluer des sujets de tous niveaux, le test devra lui procurer suffisamment d'information sur toute l'étendue du trait. Par conséquent, avant de sélectionner les items d'un test, il est nécessaire de déterminer la courbe d'information souhaitée.

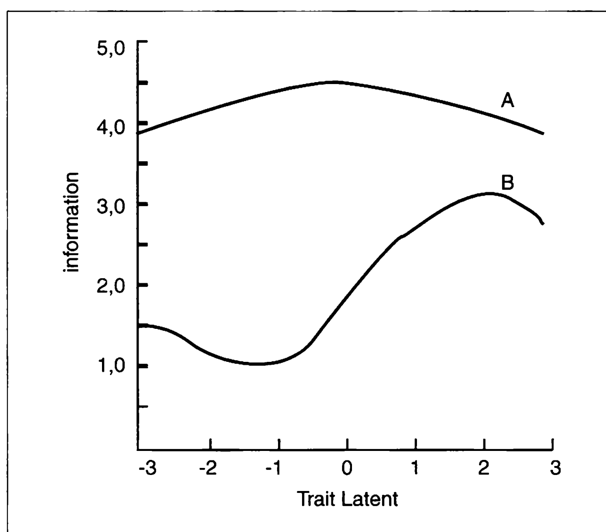


Figure 4 – Exemples de fonctions d'information de deux tests

À partir de la fonction d'information du test, nous pouvons calculer l'erreur type d'estimation du niveau d'aptitude à l'aide de la formule suivante :

$$SE(\hat{\theta}) = \frac{I}{\sqrt{I(\theta)}} \qquad (8.10)$$

Connaissant l'information du test à un point donné de θ, nous pouvons ainsi déterminer un intervalle de confiance autour de l'estimation de la capacité d'un sujet se situant à ce point θ. Comme dans la théorie classique, plus cet intervalle de confiance est étroit, plus l'estimation de l'aptitude peut être considérée comme précise. L'erreur type d'estimation dépend du nombre d'items utilisés pour estimer l'aptitude d'un sujet. Elle dépend également du pouvoir discriminatif des items et de l'adéquation de leur niveau de difficulté au niveau d'aptitude du sujet évalué. À la différence de ce que postule la théorie classique, cette erreur d'estimation peut varier selon le niveau d'habileté des sujets puisqu'elle dépend de la fonction d'information des items correspondant aux différentes valeurs de θ.

Le tableau 2 permet d'illustrer ce phénomène de variabilité de l'erreur d'estimation en fonction du niveau θ. Il présente un extrait des résultats de l'analyse des items d'un test de vocabulaire réalisée avec le logiciel RASCAL 3.5. (Assessment System Corporation, 1992). Pour chaque item, ce tableau indique le niveau de difficulté,

l'erreur type de mesure et le résultat du test d'ajustement χ^2. Les valeurs de χ^2 suivies d'une ou deux astérisques indiquent des items mal ajustés aux exigences du modèle. À la lecture de la seconde colonne, nous pouvons observer que les items varient en difficulté. Le niveau d'habileté θ nécessaire pour avoir 50% de chance de les réussir s'étend en effet de -2,803 (item 11) à 1,906 (item 22). Dans la troisième colonne, nous constatons que l'erreur type de mesure diffère sensiblement d'un item à l'autre. La précision de l'estimation de l'habileté sera affectée par ces différences d'erreur type. L'évaluation des sujets de faible capacité (θ inférieur à -1) sera moins précise que celle des sujets de niveau moyen (θ entre -1 et +1).

Tableau 2 – Analyse selon le modèle de Rasch des items d'un test de vocabulaire. Extrait des résultats (Grégoire & al., 1996)

N° item	Difficulté	Erreur type	χ^2
...	...	...	...
11	-2.803	0.254	17.489
12	-0.884	0.202	14.146
13	-1.659	0.219	23.396
14	-1.298	0.210	16.518
15	-0.766	0.200	16.634
16	-2.213	0.234	8.375
17	0.578	0.185	132.520**
18	-0.463	0.195	30.337*
19	0.276	0.187	15.350
20	0.511	0.185	12.073
21	1.413	0.187	20.133
22	1.906	0.193	12.535
23	0.578	0.185	41.004**
24	-0.390	0.194	21.537
25	0.678	0.185	39.142**
26	-0.245	0.192	13.616
...	...	...	...

5. Applications des MRI

5.1 Analyse du fonctionnement différentiel des items

Il est possible de comparer le fonctionnement d'un item dans deux sous-groupes de la population en traçant sur un même graphique les CCI de cet item pour chacun des sous-groupes en question. La figure 5 présente les courbes caractéristiques d'un même item pour deux sous-groupes de la population. Dans ce cas, le fonctionnement différentiel est uniforme, c'est-à-dire que les deux CCI ne se croisent pas. Le groupe A est en effet avantagé par cet item à tous les niveaux d'aptitude. En d'autres termes, à un même niveau d'aptitude, un sujet appartenant au groupe A aura plus de chance de réussir cet item qu'un sujet du groupe B. Sur le graphique, nous avons indiqué la différence de probabilité de réussite selon le groupe d'appartenance pour deux sujets se situant à la valeur 1 sur le trait latent.

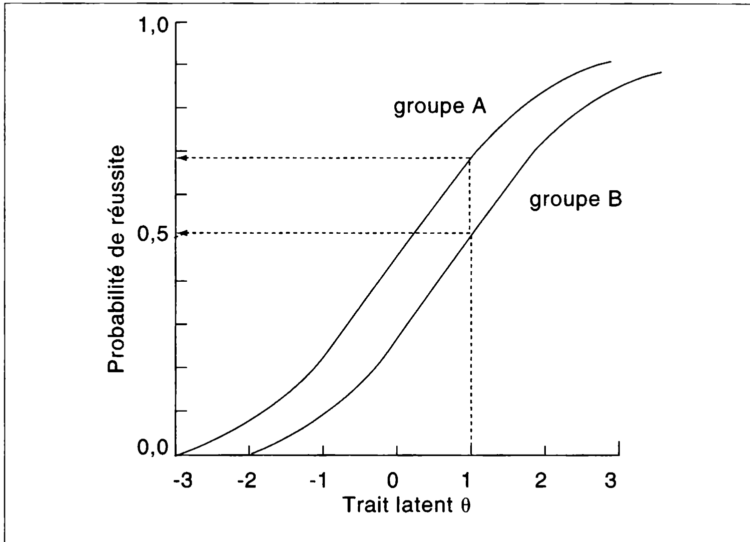


Figure 5 – Courbe caractéristique d'un même item pour deux sous-groupes de la population

Dans la figure 5, la différence entre les fonctions caractéristiques d'un même item se situe au niveau du paramètre de difficulté. Mais la différence peut aussi apparaître au niveau des paramètres de discrimination et de pseudo-chance. Ces deux cas peuvent donner lieu à un fonctionnement différentiel non uniforme. Dans ce dernier cas, l'avantage ne bénéficie pas au même groupe à tous les niveaux d'aptitude. Un item peut ainsi être plus facile pour un groupe aux niveaux d'aptitude les plus faibles alors qu'il est plus facile pour l'autre groupe à des niveaux d'aptitude plus élevés. Par rapport aux autres méthodes, les MRI sont particulièrement puissants pour repérer un fonctionnement différentiel non uniforme; pour autant que nous utilisions un modèle à deux ou à trois paramètres.

Les méthodes de détection du fonctionnement différentiel des items basées sur les MRI peuvent se ranger en deux grandes catégories : (1) celles qui utilisent les paramètres et (2) celles qui utilisent le calcul de l'aire.

L'indice le plus simple pour comparer le fonctionnement différentiel d'un item est la différence entre les estimations du paramètre de difficulté (b) entre deux groupes. Le signe de cette différence indique pour quel groupe l'item est le plus facile. Il est également possible de comparer les paramètres de discrimination (a) et de pseudo-chance (c). Toutefois, ces dernières comparaisons ne sont pas recommandées car l'estimation des paramètres a et c est généralement moins stable que celle du paramètre b. Par conséquent, ces comparaisons sont moins fiables. Elles sont d'ailleurs peu utilisées (Camilli & Shepard, 1994, p.69). Un test de signification pour la différence entre les paramètres b a été proposé par Lord (1980) :

$$d = \frac{\Delta b}{S_{\Delta b}} \qquad (8.11)$$

Δb = la différence entre les estimations de b dans les deux groupes,

$S_{\Delta b} = \sqrt{S_A^2 + S_B^2}$ = l'erreur type de la différence,

S_A et S_B sont respectivement l'erreur type d'estimation de b dans le groupe A et dans le groupe B.

Comme d se distribue à peu près normalement, la table de z peut être utilisée pour tester l'hypothèse : $H_0 : \Delta b = 0$.

Un autre groupe de méthodes de détection du fonctionnement différentiel utilise le calcul de l'aire entre les deux courbes. Rudner & al. (1980) ont proposé la formule suivante pour permettre d'évaluer l'importance de cette aire :

$$Aire = \int (P_A(\theta) - P_B(\theta))\, d\theta \tag{8.12}$$

Il existe diverses variantes de cette formule permettant de prendre en compte le fonctionnement différentiel non uniforme et l'existence d'une différence de fiabilité entre certaines portions des CCI.

5.2 Le testing adaptatif

Le principe du testing adaptatif remonte à l'origine des tests psychologiques. En effet, dès 1909, Binet proposait déjà de n'administrer aux sujets que les items les plus proches de leur niveau d'aptitude. Par exemple, à un enfant de 9 ans, on proposait d'abord les items réussis en moyenne par les enfants de cet âge. S'il les réussissait, on passait à des items plus difficiles; si au contraire il les échouait, on présentait des items plus faciles. En procédant de la sorte, on évitait de présenter à l'enfant un grand nombre d'items trop faciles ou trop difficiles. On pouvait ainsi réduire le temps de passation tout en obtenant une mesure suffisamment précise.

Le testing adaptatif est donc un testing « sur mesure » (*tailored testing*) qui a pour principal avantage le gain de temps. Il évite également que le sujet se démotive en ayant à répondre à un grand nombre d'items trop simples ou en ayant à subir une longue liste d'items qu'il échoue systématiquement. Cette démotivation peut avoir un impact sur la qualité de la mesure recueillie. Le testing adaptatif permet ainsi une meilleure mesure au moindre coût.

Toutefois, le testing adaptatif complexifie la procédure de passation. Jusqu'il y a peu, ce mode de testing n'était possible qu'en situation d'évaluation individuelle. Il fallait en effet que le praticien réalise le travail de sélection des items tout au long de la passation. L'introduction du testing sur ordinateur a permis de sortir de ce carcan. La machine peut aujourd'hui réaliser un testing sur mesure beaucoup plus ajusté aux compétences des sujets que ne le permettait la méthode non automatisée. Le sujet reçoit les items sur écran et y répond généralement à l'aide du clavier ou de la souris. En fonction de la qualité de la réponse, l'ordinateur choisit un autre item. Et ainsi de suite, jusqu'au moment où la mesure atteint le degré de précision souhaité.

Un testing adaptatif efficient n'est guère réalisable dans le cadre de la théorie classique des tests. En effet, nous avons déjà souligné que les caractéristiques métri-

ques des items sont alors relatives. Ce problème est particulièrement aigu dans le cas des tests adaptatifs car l'ordinateur doit disposer d'une vaste banque d'items où il peut sélectionner les items les plus adéquats. Or, nous avons déjà souligné plus haut que, du fait de leur nombre, tous les items d'une telle banque ne sont habituellement pas analysés avec les mêmes groupes de sujets. Par conséquent, si nous utilisons la théorie classique, les items qui composeront la banque posséderont des caractéristiques métriques relatives. Comment dès lors composer un test sur mesure avec de tels items ? Une seconde limite de la théorie classique pour le testing adaptatif vient du fait que le coefficient de fiabilité et l'erreur type de mesure sont toujours calculés pour l'entièreté du test. Si nous changeons la composition du test, l'erreur type de mesure est automatiquement modifiée. Or, le principe même du testing adaptatif est d'évaluer les sujets à partir d'ensembles d'items constitués sur mesure. Comment, dans ces conditions, déterminer l'erreur d'estimation ? Cette information est essentielle pour le testing adaptatif car la procédure consiste à réduire progressivement l'erreur d'estimation jusqu'à un point déterminé a priori.

Pour les deux raisons principales que nous venons d'examiner, il est nécessaire de se tourner vers les MRI lorsque nous voulons réaliser des tests adaptatifs. Ces modèles nous permettent en effet d'obtenir des paramètres d'items invariants. Ils nous permettent également de déterminer un intervalle de confiance pour chaque estimation du trait θ sur base de l'ensemble des items effectivement présentés au sujet.

Tableau 3 – Exemple numérique d'une procédure de testing adaptatif (d'après Urry, 1977)

Présentation	Numéro des items	Réponse	Estimation de l'aptitude	Erreur d'estimation
1	43	réussite	0,47	0,86
2	57	réussite	0,93	0,75
3	55	réussite	1,27	0,64
4	12	réussite	1,44	0,57
5	13	réussite	1,59	0,53
6	54	réussite	1,77	0,50
7	114	réussite	1,88	0,47
8	26	réussite	1,98	0,43
9	103	*échec*	1,80	0,39
10	79	réussite	1,87	0,38
11	78	réussite	1,95	0,37
12	149	*échec*	1,80	0,34
13	15	réussite	1,85	0,33
14	76	réussite	1,88	0,32
15	74	réussite	1,94	0,32

Dans les programme de testing adaptatif, l'algorithme utilisé prend en compte la fonction d'information de chaque item (Thissen & Mislevy, 1990). Ainsi, à chaque fois que le sujet a répondu à un item, l'ordinateur réestime son aptitude et recalcule l'erreur type de cette estimation. Sur base de l'estimation obtenue et de sa marge d'erreur, l'ordinateur peut alors choisir l'item qui procurera la plus grande quantité d'information au niveau θ considéré. Habituellement, la procédure commence par un item de difficulté moyenne. À partir de la réussite ou de l'échec à ce premier item, un second item est choisi. Et ainsi de suite. Au cours de cette procédure, l'aptitude du sujet est systématiquement réestimée ainsi que l'erreur d'estimation. Le testing s'arrête lorsque l'on atteint un niveau d'erreur spécifié à l'avance. Le tableau 3 nous présente une illustration de cette procédure. On peut constater qu'à partir d'un moment, l'estimation de l'aptitude et la marge d'erreur se stabilisent. Il n'y a dès lors plus lieu de continuer la procédure d'évaluation.

Les tests adaptatifs nous permettent d'obtenir une mesure très précise avec un minimum d'items. Toutefois, il faut avoir conscience que pour réaliser une telle évaluation, une banque importante d'items est nécessaire. Il n'est pas rare que cette banque comprenne plusieurs centaines d'items alors que chaque sujet individuellement n'en voit qu'une quinzaine.

6. Quel MRI choisir?

Quel modèle choisir parmi les trois modèles de réponse à l'item que nous avons présentés ? En ce moment, le modèle de Rasch semble être le plus couramment utilisé. Un des arguments qui joue le plus en faveur de ce modèle est la taille relativement réduite de l'échantillon de sujets nécessaire pour obtenir une estimation correcte du paramètre de difficulté. Toutefois, nous ne devons pas perdre de vue que ce modèle repose sur le postulat que tous les items sont également discriminatifs. Ce postulat peut conduire à écarter un grand nombre d'items mal ajustés au modèle. Comme le fait remarquer Hambleton (1994), dans un cas comme celui-là, il est légitime de se demander si ce n'est pas le modèle qui doit être remis en question. En éliminant les items dont le degré d'adéquation au modèle de Rasch est insuffisant, nous risquons en effet de nous priver de certains de nos items les plus valides. Dans ce cas, il est raisonnable de vérifier si les modèles à deux ou à trois paramètres ne sont pas plus adaptés pour nos données que le modèle à un seul paramètre. En fait, le choix du modèle doit nous permettre d'obtenir un meilleur ajustement de nos données et, par là même, une estimation plus précise et plus stable des paramètres des items.

Les logiciels actuels, qu'ils soient basés sur un modèle à un, deux ou trois paramètres, ont tous été conçus pour réaliser des analyses d'items dichotomiques, c'est-à-dire d'items cotés 1 ou 0. Or, les praticiens ont souvent affaire à des items multichotomiques. Par exemple, de nombreux questionnaires demandent de répondre sur une échelle de 1 à 5. Les logiciels comme BILOG ou XCALIBRE ne permettent pas de traiter de telles données. Par conséquent, si nous désirons réaliser une analyse selon un des MRI, nous devons tout d'abord dichotomiser nos items. Ceci entraîne une perte d'information et soulève des questions de validité parfois difficiles à résoudre. Par exemple, lorsqu'un item est coté 0, 1 ou 2, vaut-il mieux regrouper les résultats 0 et 1 ou les résultats 1 et 2 ? Des recherches sont en cours visant à mettre au point des logi-

ciels pour traiter des données multichotomiques dans le cadre des MRI. Certains sont déjà à la disposition des praticiens comme MULTILOG (Thissen, 1986) et TESGRAF (Ramsay, 1991 et 1993).

Un autre problème soulevé par les MRI actuels concerne le postulat d'unidimensionnalité. Nous avons déjà souligné que, si nos donnés ne satisfont pas ce postulat, l'utilisation d'un des MRI que nous avons présentés plus haut n'a guère de sens. Or, en psychologie et en éducation, les performances à de nombreux tests sont déterminées par plusieurs facteurs sous-jacents, indépendants ou corrélés. Dans ce cas, le postulat d'unidimensionnalité n'est pas défendable et l'analyse des items selon un des MRI unidimensionnels n'est pas possible. Plusieurs chercheurs ont proposé des MRI multidimensionnels (par exemple, Embretson, 1991) mais aucun n'a encore débouché sur des applications pratiques concluantes. Cette carence vient sans doute du fait *« qu'une grande part des développements mathématiques durant ces cinquante dernières années se sont concentrées trop exclusivement sur le cas particulier des modèles logistiques unidimensionnels »* (Goldstein & Wood, 1989, p.164).

BIBLIOGRAPHIE

AMERICAN PSYCHIATRIC ASSOCIATION (1994). *Diagnostic and statistical manual of mental disorders. DSM-IV.* Washington, DC : American Psychiatric Association.

AMERICAN PSYCHOLOGICAL ASSOCIATION (1985). *Standards for educational and psychological testing*. Washington, DC : American Psychological Association.

AMERICAN PSYCHOLOGICAL ASSOCIATION (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC : American Psychological Association.

ANASTASI, A. (1982). *Psychological testing (5th ed.).* New York : Mcmillan.

ANGOFF, W.H. (1971). Scales, norms and equivalent scores. In R.L.THORNDIKE (Ed.), *Educational measurement.* Washington : American Council on Education.

ANGOFF, W.H. (1988). Validity : An evolving concept. IN H. WAINER & H.I. BRAUN (Eds.), *Test validity.* Hillsdale, NJ : Lawrence Erlbaum.

ANGOFF, W.H. (1993). Perspectives on differential item functioning methodology. In P. Holland & H. Wainer (Eds.). *Differential item functioning.* Hillsdale, NJ : Lawrence Erlbaum.

ASSESSMENT SYSTEMS CORPORATION. (1992). *Rascal, version 3.5.* St. Paul, MN : Assessment Systems Corporation.

ASSESSMENT SYSTEMS CORPORATION. (1994). *Xcalibre, version 1.0.* St. Paul, MN : Assessment Systems Corporation.

BARKER, D. & EBEL, R.L. (1981). A comparison of difficulty and discrimination values of selected true-false item types. *Contemporary Educational Psychology, 7*, 35-40.

BEUCHERT, A.K. & MENDOZA, J.L. (1979). A Monte Carlo comparison of ten item discrimination indices. *Journal of Educational Measurement, 16*, 109-118.

BIRNBAUM, A. (1968). Somme latent trait models and their use in inferring an examinee's ability. In F.M. LORD & M.R.NOVICK (Eds.), *Statistical theories of mental tets scores.* Reading, MA : Addison-Wesley.

BLOOM, B.S. (Ed.). (1956). *Taxonomy of educational objectives : Handbook I, cognitive domaine*. New York : D. McKay.

BLOOM, B.S. (1984). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership, 5*, 4-17.

BRENNAN, R.L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement, 32*, 289-303.

CAMILLI, G. & SHEPARD, L.A. (1994). *Methods for identifying biased test items.* London : Sage.

CAMPBELL, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

CARDINET, J. & TOURNEUR, Y. (1985). *Assurer la mesure.* Berne : Peter Lang.

CARDINET, J. (1987). *Évaluation des élèves et pédagogie active.* Neuchâtel : Institut Romand de Recherches et de Documentation Pédagogiques.

COLE, N.C. & MOSS, P.A. (1989). Bias in test use. IN R. LINN (Ed.). *Educational Measurement.* New York : American Council on Education & Macmillan.

CONSEIL SUPÉRIEUR DE L'EDUCATION (1992). *Évaluer les apprentissages au primaire : un équilibre à trouver.* Québec : Direction des Communications, Conseil Supérieur de l'Education.

COX, R.C. & VARGAS, J.S. (1966). *A comparison of item selection techniques for norm-referenced and criterion-referenced tests.* Paper presented at the annual meeting of the National Council on Measurement in Education.

CROCKER, L. & ALGINA, J. (1986). *Introduction to classical and modern test theory.* New York : Holt, Rinehart and Winston.

CRONBACH, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-234.

CRONBACH, L.J., GLESER, G.C., NANDA, H., & RAJARATNAM, N. (1972). *The dependability of behavioral measurements.* New York : John Wiley.

CRONBACH, L.J., GLESER, G.C. & RAJARATNAM, N. (1963). Theory of generalizability. A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology, 16,* 137-173.

DANE, F.C. (1990). *Research methods.* Pacific Grove, CA : Brooks/Cole.

DE LANDSHEERE, V. (1986). *Faire réussir, faire échouer.* La compétence minimale et son évaluation. Paris : Puf.

DE PARTZ, M.-P. (1994).L'évaluation de la lecture en neuropsychologie. IN J. GRÉGOIRE & B. PIÉRART (Eds.), *Evaluer les troubles de la lecture.* Bruxelles : De Boeck.

DECHEF, H. & LAVEAULT, D. (1993). Etude du fonctionnement différentiel des items à l'aide des méthodes du khi carré, de Mantel-Haenszel et logit. *Mesure et Evaluation en Education, 16,* 5-28.

DENO, S.L. & JENKINS, J.R. (1969). On the " behaviorality " of behavioral objectives. *Psychology in the school,* 69, 18-24.

DILLON, J.T. (1984). The classification of research questions. *Review of Educational Research, 54,* 327-361.

DIXON, W.J., BROWN, M.B., ENGELMAN, L., FRANE, J.W., HILL, M.A. JENNRICH, R.I., & TOPOREK, J.D. (1981). *BMDP statistical software.* Berkeley : University of California Press.

DORANS, N.J. (1989). Two new approches to assessing differential item functioning : Standardization and the Mantel-Haenszel. *Applied Psychological Measurement, 2,* 217-233.

DYER, H.S. (1967). The discovery and development of educational goals. In J.C. STANLEY (Ed.). *Proceedings of the 1966 invitational conference on testing problems.* Princeton, N.J. : Educational Testing Service.

EBEL, R.L. & FRISBIE, D.A. (1991). *Essential of educational measurement.* Englewood Cliffs, NJ : Prentice Hall.

EBEL, R.L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement, 16,* 269-282.

EBEL, R.L. (1965). *Measuring educational achievement.* Englewood Cliffs, N.J. : Prentice-Hall.

EMBRETSON, S.E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika,* 56, 495-515.

ENGLEHART, M.D. (1965). A comparison of several item discrimination indices. *Journal of Educational Measurement, 2,* 69-76.

EVANS, J. (1968). Behavioral objectives are no dam good. In *Technology And Innovation in Education.* N.Y., 43.

EXNER, J.E. & EXNER, D.E. (1972). How clinicians use the Rorschach. *Journal of Personnality Assessment, 36*, 403-408.

EXNER, J.E. (1974). *The Rorschach : A comprensive system.* New York : Wiley.

FELT, L.S. (1984). Some relationships between the binomial error model and classical test theory. *Educational and Psychological Measurement, 44*, 883-891.

FELT, L.S., STEFFEN, M. & GUPTA, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score level. *Applied Psychological Measurement, 9*, 351-361.

FINDLEY, W. G. (1956). A rationale for evaluation of item discrimination statistics. *Educational and Psychological Measurement, 16*, 175-180.

FLANAGAN, J.C. (1954). The critical incident technique. *Psychological Bulletin Psychology, 51*, 327-358.

FLAUGHER, R.L. (1978). The many definitions of test bias. *American Psychologist, 33*, 671-679.

FLYNN, J.R. (1987). Massive gain in 14 nations : What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

FOURNIER, C. & LAVEAULT, D. (1992). L'examen de rendement scolaire : lieu de rencontre des attentes du Maître et des stratégies d'étude de l'élève. In D. LAVEAULT (Éd.). *Les pratiques d'évaluation en éducation.* Montréal : Éditions de l'ADMEE.

GALLUP, G. (1947). The quintamentional plan of question design. *Public Opinion Quarterly, 11*, 385-393.

GOLDSTEIN, H. & WOOD, R. (1989). Five decades of response modeling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

GRÉGOIRE, J. (1992). *Évaluer l'intelligence de l'enfant.* Liège : Mardaga.

GRÉGOIRE, J. (1994). Application de la méthode de Mantel-Haenszel à l'analyse du fonctionnement différentiel des items du K-ABC entre filles et garçons. *Revue Européenne de Psychologie Appliquée, 45*, 111-118.

GRÉGOIRE, J., PENOUET, C. & BOY, Th. (1996). L'adaptation française de l'échelle de Wechsler pour enfants, version III. *L'Orientation Scolaire et Professionnelle, 25*, 489-506.

GORSUCH, R.L. (1983). *Factor analysis.* Philadelphia : Saunders.

GRONLUND, N.E. (1991). *How to construct achievement tests.* Needham Heights, MA : Allyn and Bacon.

GUILFORD, J.P. (1954). *Psychometric methods.* New York : McGraw-Hill.

GULLIKSEN, H. (1950). *Theory of mental tests.* New York : Wiley.

GUTTMAN, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*, 255-282.

GUTTMAN, L. (1950). The basis for scalogram analysis. In S.A. STOUFFER (Ed.), *Measurement and prediction.* Princeton, NJ : Princeton University Press.

GUTTMAN, L. (1969). Integration of test design and analysis. Proceding of the 1969 invitation conference on testing problems. Princeton, NJ : Princeton University Press.

HAMBLETON, R.K. (1980). Test score validity and standard setting methods. In R.A. BERK (Ed.). *Criterion-referenced measurement : The state of the art.* Baltimore : Johns Hopkins University Press.

HAMBLETON, R.K. (1994). Item response theory : a broad psychometric framework for measurement advances. *Psicothema, 6*, 535-556.

HAMBLETON, R.K. & JONES, R. (1993). Comparison of empirical and judgemental procedures for detecting differential item functioning. *Educational Research Quarterly.*

HAMBLETON, R.K. & MURRAY, L.N. (1983). Some goodness of fit investigations for item response models. In R.K. HAMBLETON (Ed.), *Applications of item response theory.* Vancouver, Bs : Educational Research Institute of British Columbia.

HAMBLETON, R.K. & ROGERS, H.J. (1989). Detecting potentially biaised test items : comparaison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2*, 313-334.

HAMBLETON, R.K. & SWAMINATHAN, H. (1985). *Item response theory.* Principles and applications. Norwell, MA : Kluwer.

HAMBLETON, R.K., CLAUSER, B.E., MAZOR, K.M. & JONES, R.W. (1993). Advances in detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*, 1-18.

HAMBLETON, R.K., SWAMINATHAN, H. & ROGERS, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA : Sage.

HARRIS, C.W. & PEARLMAN, A.P. (1977). *Conventional significance tests and indices of agreement or study.* (CSE monograph serie in evaluation, no. 6). Los Angeles : Center for the Study of Evaluation, University of California.

HARROW, A.J. (Ed.). (1972). *A taxonomy of the psychomotor domain.* New York : D. MCKAY.

HAYNES, S.N., RICHARD, D.C.S., KUBANY & E.S.(1995). Content validity in psychological assessment : A functional approach to concepts and methods. *Psychological Assessment, 7*, 238-247.

HOLLAND, P.W. & THAYER, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. WAINER & H. BRAUN (Eds.). *Test Validity.* Hillsdale, NJ : Lawrence Erlbaum Associates.

HOTELLING, H. & PABST, M.R. (1936). Rank correlation and tests of significance involving no assumption of normality. *Ann. Math. Statist., 7*, 29-43.

HOWELL, D.C. (1992). *Statistical methods for psychology.* Belmont, CA : Duxbury Press.

Hoyt, C.J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153-160.

HULIN, C.L., LISSAK, R.I. & DRASGOW, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves : a Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

HUNT, S.M. & MCKENNA, S.P. (1992). The QLDS : A scale for measurement of quality of life in depression. *Health Policy, 22*, 307-319.

JAEGER, R.M. (1989). Certification of student competence. In R. LINN (Ed.). *Educational measurement.* New York : American Council on Education/MacMillan.

JÖRESKOG, K.G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 183-202.

KANE, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425-461.

KAUFMAN, A.S. (1975). Factor analysis fo the WISC-R at 11 age levels between 6 1/2 and 16 1/2. *Journal of Consulting and Clinical Psychology, 43*, 135-147.

KEATS, J.A. & LORD, F.M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27*, 215-231.

KEATS, J.A. (1957). Estimation of error variances of test scores. *Psychometrika, 22*, 29-41.

KELLEY, T.L. (1939). Selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30*, 17-24.

KENDALL, M.G. (1938). A new measure of rank correlation. *Biometrika, 30*, 81-93.

KENDALL, M.G. (1948). *Rank correlation methods.* London : Griffin.

KLEIN, S.P. & KOSECOFF, J.P. (1975). *Determining how well a test measures your objectives.* (CSE Report No. 94). Los Angeles : Center for the Study of Evaluation, University of California.

KLOPFER, W.G. & TAULBEE, E.S. (1976). Projective tests. *Annual Review of Psychology, 27*, 543-567.

KRATHWOHL, D.R. (Ed.). (1964). *Taxonomy of educational objectives* : Handbook II, affective domaine. New York : D. McKay.

KUDER, G.F. & RICHARDSON, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

KURTZ, A.K. & MAYO, S.T. (1979). *Statistical methods in education and psychology.* New York : Springer-Verlag.

LAVEAULT, D. & FOURNIER, C. (1990). Évaluation par objectifs : une approche métacognitive. *Mesure et Évaluation en Éducation, 13*, 57-74.

LECLERCQ, D. (1986).*La conception des questions à choix multiple.* Bruxelles : Labor.

LEGENDRE, R. (1993). *Dictionnaire actuel de l'éducation* (2ème édition). Montréal : Guérin.

LIKERT, R. (1932). A technique for the measurement of attitudes.*Archives of Psychology,* 1-55.

LIVINGSTONE, S.A. & ZIELSKY, M.J. (1982). Passing scores : *A manual for setting standards on educational and occupational tests.* Princeton : Educational Testing Service.

LORD, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

LORD, F. M. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, Mass. : Addison-Wesley.

LORD, F.M. (1953a). An application of confidence intervals and maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*, 57-75.

LORD, F.M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-548.

LORD, F.M. (1953c). On the statistical treatment of numbers. *American Psychologist, 8*, 750-751.

LORD, F.M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325-326.

LORD, F.M. (1959). Test norms and sampling theory. *Journal of Experimental Education, 27*, 247-263.

LORD, F.M. (1965). A strong true-score theory with application. *Psychometrika, 30*, 239-270.

LORD, F.M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ : Lawrence Erlbaum.

MAGNUSSON, D. (1967). *Test theory.* Boston : Addison-Wesley.

MATALON, B. (1965). *L'analyse hiérarchique.* Paris : Gauthier-Villars.

MAZOR, K.M., CLAUSER, B.E. & HAMBLETON, R.K. (1992). The effect of sample size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*, 443-451.

MCDONALD, R. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.

MESSICK, S. (1988). The once and the future issues of validity : Assessing the meaning and consequences of measurement. In H. WAINER & H.I. BRAUN (Eds.), *Test validity.* Hillsdale, NJ : Lawrence Erlbaum.

MESSICK, S. (1989). Validity. In R.L.Linn, *Educational measurement.* Washington : American Council on Education/McMillan.

MESSICK, S. (1995). Validity of psychological assessment.*American Psychologist, 50*, 741-749.

MISLEVY, R.J. & BOCK, R.D. (1990). *Bilog, Version 3.* Mooresville, IN : Scientific Software.

MORRIS, C.N. (1982). On the foundation of test equating. In P.W. HOLLAND & D.B. RUBIN (Eds.) *Test equating.* New York : Academic Press.

MOUSTY, PH., LEYBAERT, J., ALÉGRIA, J. DELTOUR, J.-J. & SKINKEL, R. (1994). BELEC, une batterie d'évaluation du langage écrit et de ses troubles. In J. GRÉGOIRE & B. PIÉRARt (Eds.), *Evaluer les troubles de la lecture.* Bruxelles : De Boeck.

NEDELSKY, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3-19.

NANDAKUMAR, R., GLUTTING, J.J. & OAKLAND, T. (1993). Mantel-Haenszel methodology for detecting item bias. *Journal of Psychoeducational Assessment, 11*, 108-119.

OOSTERHOF, A.C. (1976). Similarity of various item discrimination indices. *Journal of Educational Measurement, 13*, 145-150.

OSTERLIND, S.J. (1989). *Test item bias*. Newbury Park, CA : Sage.

PAQUAY, L., ALLAL, L. & LAVEAULT, D. (1990). L'autoévaluation en question(s). Propos pour un débat. *Mesure et Évaluation en Éducation, 13*, 5-26

PETERSEN, N.S., KOLEN, M.J. & HOOVER, H.D. (1988). Scaling, norming and equating. In R.L. LINN (Ed.), *Educational measurement.* New York : American Council on Education/MacMillan.

POPHAM, W.J. (1980). Domain specification strategies. In R. A. BERK (Ed.). *Criterion-referenced measurement, the state of the art.* Baltimore : The John Hopkins University Press.

RAMSAY, J. (1991). Kernel smoothing approaches to non parametric item characteristic curve estimation. *Psychometrika, 56*, 611-630.

RAMSAY, J. (1993). *Tesgraf. A program for graphical analysis of multiple choice test and questionnaire data.* Montréal : McGill University.

RESCHLY, D.J. (1978). WISC-R factor structures among anglos, blacks, chicanos and native-american papagos. *Journal of Consulting and Clinical Psychology, 46*, 417-422.

ROID, G.H. & HALADYNA, T.M. (1982). *A technology for test-item writing.* New York : Academic Press.

RUDNER, L.M., GETSON, P.R. & KNIGHT, D.L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213-233.

RULON, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99-103.

SATTLER, J.M. (1988). *Assessment of children..* San Diego : Jerome M. Sattler Publisher.

SCALLON, G. (1992). L'évaluation formative : entre la docimologie et la didactique. In D. Laveault (Ed.). *Les pratiques d'évaluation en éducation.* Montréal : Éditions de l'ADMEE.

SCHEUNEMAN, J.C. & BLEINSTEIN, C.A. (1989). A consumer's guide to statistics for differential item functioning. *Applied Measurement in Education, 2*, 255-275.

SCHMIDT, F.L., HUNTER, J.E. & URRY, V.W. (1976). Statistical power in criterion-related validity studies. *Journal of Applied Psychology, 61*, 473-485.

SECOLSKY, C. (1983). Using examinee judgments for detecting invalid items on teacher-made criterion-referenced tests, *Journal of Educational Measurement, 20*, 51-63.

SHOEMAKER, D.M. (1975). Toward a framework for achievement testing. *Review of Educational Research, 45*, 127-148.

SPEARMAN, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

SIEGEL, S. & CASTELLAN, N.J. (1988). *Nonparametric Statistics for the Behavioral Sciences* (2nd edition). New York : McGraw-Hill.

STEVENS, S. S. (1946). On the theory of scales of measurement. *Science, 103*, 677-680.

THISSEN, D. M. & MISLEVY, D. (1990). Testing algorithms. In H. WAINER (Ed.), *Computerized adaptative testing.* A primer. Hillsdale, NJ : Lawrence Erlbaum.

THISSEN, D. M.(1986). *Multilog : Item analysis and scoring with multiple category response model.* Moresville, IN : Scientific Software.

THORNDIKE, R.L. (1949). *Personnel selection : test and measurement techniques.* New York : Wiley.

THORNDIKE, R.L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8*, 63-70.

THURSTONE, L.L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

TUKEY, J.W. (1977). *Exploratory data anlysis.* Reading, MA : Addison-Wesley.

URRY, V.W. (1977). Tailord testing : A successful application of item response theory. *Journal of Educational Measurement, 14*, 181-196.

VALE, C.D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*, 333-344.

VAN DER LINDEN, W. (1986). The changing conception of measurement in education and psychology. *Applied Psychological Measurement, 10*, 325-332.

WALLBROWN, F.H., BLAHA, J., WALLBROWN, J. & ENGIN, A. (1975). The hierarchical factor structure of the Wechsler intelligence scale for children-revised. *Journal of Psychology, 89*, 223-235.

WECHSLER, D. (1981). *Manuel de l'échelle de Wechsler pour enfants, forme révisée.* Paris : Editions du Centre de Psychologie Appliquée.

WIERSMA, W. & JURS, S.G. (1990). *Educational Measurement and Testing.* Boston : Allyn and Bacon.

WRIGHT, B.D. & STONE, M.H. (1979). *Best test design.* Chicago : Mesa Press.

GLOSSAIRES

GLOSSAIRE DES TERMES TECHNIQUES ET TRADUCTION FRANCAISE

Anglais	**Français**
Achievement test	*Test de connaissances* *Test d'acquisition(s)*
Adaptive testing	*Testing adaptatif*
Age-equivalent score	*Score en niveau d'âge*
Alternate choice item	*Question " vrai-faux "*
Anchor test	*Test d'ancrage*
Aptitude test	*Test d'aptitude*
Assessment	*Évaluation*
Bias	*Biais*
Binomial distribution	*Distribution binomiale*
Binomial error model	*Modèle binomial de l'erreur*
Clue	*Indice*
Cluster sampling	*Échantillonnage par grappes*
Coefficient of determination	*Coefficient de détermination*
Composite score	*Score composite*
Compound binomial error model	*Modèle binomial composite de l'erreur*
Concurrent validity	*Validité concomitante*
Confidence interval	*Intervalle de confiance*
Construct	*Concept hypothétique* *Modèle*
Construct validity	*Validité conceptuelle* *Validité théorique*
Constructed-response item	*Item à réponse construite*
Content validity	*Validité de contenu*
Convergent validity	*Validité convergente*
Covariance	*Covariance*
Criterion	*Critère*
Criterion validity	*Validité critérielle* *Validité liée à un critère*

Criterion-referenced test	*Test critérié*
Critical incident	*Incident critique*
Culture-fair test	*Test culturellement équitable*
Cut score	*Note de césure*
	Score seuil
Decision consistency	*Cohérence de la décision*
Degree of Freedom	*Degré de liberté*
Design	*Plan*
	Dispositif
Dichotomous	*Dichotomique*
Differential Item Functioning	*Fonctionnement différentiel d'item*
Difficulty index	*Indice de difficulté*
Dimension	*Dimension*
Distractor	*Leurre*
	Distracteur
Equate	*Apparier*
Equating	*Mise en équivalence*
Equipercentile equating	*Mise en équivalence équipercentile*
Error variance	*Variance d'erreur*
Essay item	*Question à réponse narrative*
Expected frequency	*Fréquence théorique*
Expected value	*Valeur attendue*
	Valeur théorique
Extended response	*Réponse étendue*
Facet	*Facette*
Factor Analysis	*Analyse factorielle*
Factor loading	*Saturation factorielle*
Field-test	*Prétest*
Focal group	*Groupe focal*
Forced-choice item	*Item à choix forcé*
Goodness-of-fit	*Ajustement*
Grade-equivalent score	*Score en niveau scolaire*
Guessing	*Choix aléatoire*
	Choix au hasard
Guttman scale	*Échelle de Guttman*
Information function	*Fonction d'information*
Internal consistency	*Cohérence interne*
Item characteristic curve	*Courbe caractéristique de l'item*
Item characteristic function	*Fonction caractéristique de l'item*
Item difficulty	*Difficulté de l'item*
Item response model (IRM)	*Modèle de la réponse à l'item*
Kurtosis	*Kurtose*
Latent trait (θ)	*Trait latent*(θ)
Latent trait theory	*Théorie des traits latents*
Least squares method (LSM)	*Méthode des moindres carrés*
Linear equating	*Mise en équivalence linéaire*
Linking	*Liaison*

Local independence	*Indépendance locale*
Main effect	*Effet principal*
Mastery test	*Test de maîtrise*
Matching item	*Question d'appariement*
Maximum likelihood estimation (MLE)	*Estimation du maximum de vraisemblance*
Meansquare (MS)	*Carré moyen (CM)*
Measurement	*Mesure*
Measurement error	*Erreur de mesure*
Measurement of change	*Mesure du changement*
Moments	*Moments*
Monotone	*Monotone*
Multidimensional	*Multidimensionnel*
Multiple-choice item	*Question à choix multiple*
Norm-referenced test	*Test normé*
Normal distribution	*Distribution normale*
Normalized z-score	*Score z normalisé*
Observed frequency	*Fréquence observée*
One-parameter logistic model	*Modèle logistique à 1 paramètre*
Partition of the variance	*Répartition de la variance*
Percentage of agreement	*Pourcentage d'accord*
Percentile rank	*Rang centile*
Performance item	*Question de performance*
Polychotomous	*Polychotomique*
Predictive validity	*Validité prédictive*
Premise	*Prémisse*
Principle component analysis	*Analyse en composantes principales*
Pseudo-guessing parameter	*Paramètre de pseudo-chance*
Quota sampling	*Échantillonnage par quota*
Random	*Aléatoire*
Rating scale	*Échelle de cotation*
Raw score	*Score brute*
Reference group	*Groupe de référence*
Reliability	*Fiabilité*
Reliability coefficient	*Coefficient de fiabilité*
Residual	*Résidu*
Restricted response	*Réponse contrainte*
Sampling	*Échantillonnage*
Scaled score	*Score d'échelle*
Scaled test	*Test gradué*
Scaling	*Échelonnage*
Scatterplot	*Graphique de dispersion*
Scoring	*Cotation*
Short-answer item	*Question à réponse brève*
Skewness	*Asymétrie*
Slope	*Pente*
Smooth curve	*Courbe lissée*

Smoothing	*Lissage*
Standard	*Standard*
Standard deviation	*Écart type*
Standard error	*Erreur type*
Standard error of estimate	*Erreur type d'estimation*
Standard error of measurement	*Erreur type de mesure*
Standardization	*Étalonnage*
Standardize	*Étalonner* *Standardiser*
Standardized test	*Test standardisé*
Stanine	*Stanine*
Stanine scale	*Échelle en stanine*
Stratified sampling	*Échantillonnage stratifié*
Subscore	*Sous-score*
Sum of squares (SS)	*Somme des carrés (SC)*
Systematic sampling	*Échantillonnage systématique*
Tailored testing	*Testing sur mesure*
Taxonomy	*Taxonomie*
Test characteristic curve	*Courbe caractéristique du test*
Test specification	*Spécification d'un test*
Testing	*Testing*
Tetrachoric correlation	*Corrélation tétrachorique*
True score	*Score vrai*
True-false item	*Item à réponse " vrai-faux "*
Unidimensional	*Unidimensionnel*
Validity	*Validité*
Variance	*Variance*
Variance-covariance matrix	*Matrice des variances-covariances*
Weighted score	*Score pondéré*

GLOSSAIRE DES PRINCIPAUX SYMBOLES

A	asymétrie
a_i	paramètre de discrimination de l'item (MRI)
α	lettre grecque *alpha* en minuscule
α	niveau de signification (erreur de type I)
α	coefficient de cohérence interne (alpha de Cronbach)
b_i	paramètre de difficulté de l'item (MRI)
β	lettre grecque *bêta* en minuscule
β	erreur de type II
c_i	coefficient de pseudo-chance (MRI)
CCI	courbe caractéristique d'item

χ	lettre grecque *chi* en minuscule
χ^2	chi-carré
d_i	indice de discrimination de Findley
dl	degré de liberté
E	espérance mathématique
e	constante de Neper ≅ 2.718
F	rapport de Fisher (ANOVA)
f(x)	fonction de x
ϕ	lettre grecque *phi* en minuscule
ϕ	coefficient de corrélation *phi*
FDI	fonctionnement différentiel d'item
I	Intervalle semi-interquartile
K	kurtose
KR20	coefficient de Kuder-Richardson, formule 20
KR21	coefficient de Kuder-Richardson, formule 21
MRI	modèle de réponse à l'item
m	moyenne de l'échantillon
μ	lettre grecque *mu* minuscule
μ	moyenne de la population
Md	Médiane
N	taille de la population
n	taille de l'échantillon
P_i	coefficient de difficulté de i
r	corrélation de Pearson (échantillon)
ρ	lettre grecque *rho* en minuscule
ρ	corrélation de Pearson (population)
s	écart type de l'échantillon
σ	lettre grecque *sigma* en minuscule
σ	écart type de la population
S_E	erreur type de l'échantillon
σ_E	erreur type de la population
s^2	variance de l'échantillon
$s^2_{\dot{Y}X}$	erreur type d'estimation
$S^2_{\Delta E}$	erreur type de la différence
σ^2	variance de la population

s_{XY}^2	covariance de l'échantillon
σ_{XY}^2	covariance de la population
t	t de Student (comparaison de moyennes)
TCS	théorie classique des scores
θ	lettre grecque thêta en minuscule
θ	variable latente du niveau d'habileté (MRI)
X	score ou variable indépendante
Y	score ou variable dépendante
$\hat{Y}$	valeur prédite de Y
$\bar{X}$	moyenne des X
z	score centré réduit ou score standard

OPÉRATEURS

$\|a\|$	valeur absolue de a
Σ	lettre grecque sigma en majuscule
Σ	sommation de toutes les valeurs
Π	lettre grecque pi en majuscule
Π	multiplication de toutes les valeurs
$<, \leq$	plus petit, plus petit ou égal
$>, \geq$	plus grand, plus grand ou égal
$\cong$	approximativement égal à
$\neq$	différent, inégal
∞	infini

INDEX DES SUJETS

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

INDEX DES NOMS D'AUTEURS

A

B

C

D

E

F

G

TABLE DES MATIÈRES

CHAPITRE 5

CHAPITRE 6